AMERICAN SOCIETY FOR MICROBIOLOGY | mSystems®

RESEARCH ARTICLE
Clinical Science and Epidemiology

Using DNA Metabarcoding To Evaluate the Plant Component of Human Diets: a Proof of Concept

Aspen T. Reese,[a] Tyler R. Kartzinel,[b] Brianna L. Petrone,[c,d] Peter J. Turnbaugh,[e] Robert M. Pringle,[f] Lawrence A. David[d,g]

[a]Society of Fellows, Harvard University, Cambridge, Massachusetts, USA
[b]Department of Ecology and Evolutionary Biology, Brown University, Providence, Rhode Island, USA
[c]Medical Scientist Training Program, Duke University, Durham, North Carolina, USA
[d]Department of Molecular Genetics and Microbiology, Duke University, Durham, North Carolina, USA
[e]Department of Microbiology and Immunology, University of California San Francisco, San Francisco, California, USA
[f]Department of Ecology and Evolutionary Biology, Princeton University, Princeton, New Jersey, USA
[g]Center for Genomic and Computational Biology, Duke University, Durham, North Carolina, USA

ABSTRACT Dietary intake is difficult to measure reliably in humans because approaches typically rely on self-reporting, which can be incomplete and biased. In field studies of animals, DNA sequencing-based approaches such as metabarcoding have been developed to characterize diets, but such approaches have not previously been widely applied to humans. Here, we present data derived from sequencing of a chloroplast DNA marker (the P6 loop of the *trn*L [UAA] intron) in stool samples collected from 11 individuals consuming both controlled and freely selected diets. The DNA metabarcoding strategy resulted in successful PCR amplification in about 50% of samples, which increased to a 70% success rate in samples from individuals eating a controlled plant-rich diet. Detection of plant taxa among sequenced samples yielded a recall of 0.86 and a precision of 0.55 compared to a written diet record during controlled feeding of plant-based foods. The majority of sequenced plant DNA matched common human food plants, including grains, vegetables, fruits, and herbs prepared both cooked and uncooked. Moreover, DNA metabarcoding data were sufficient to distinguish between baseline and treatment diet arms of the study. Still, the relatively high PCR failure rate and an inability to distinguish some dietary plants at the sequence level using the *trn*L-P6 marker suggest that future methodological refinements are necessary. Overall, our results suggest that DNA metabarcoding provides a promising new method for tracking human plant intake and that similar approaches could be used to characterize the animal and fungal components of our omnivorous diets.

IMPORTANCE Current methods for capturing human dietary patterns typically rely on individual recall and as such are subject to the limitations of human memory. DNA sequencing-based approaches, frequently used for profiling nonhuman diets, do not suffer from the same limitations. Here, we used metabarcoding to broadly characterize the plant portion of human diets for the first time. The majority of sequences corresponded to known human foods, including all but one foodstuff included in an experimental plant-rich diet. Metabarcoding could distinguish between experimental diets and matched individual diet records from controlled settings with high accuracy. Because this method is independent of survey language and timing, it could also be applied to geographically and culturally disparate human populations, as well as in retrospective studies involving banked human stool.

KEYWORDS human diet, DNA metabarcoding, *trn*L(UAA)-P6, diet log

Citation Reese AT, Kartzinel TR, Petrone BL, Turnbaugh PJ, Pringle RM, David LA. 2019. Using DNA metabarcoding to evaluate the plant component of human diets: a proof of concept. mSystems 4:e00458-19. https://doi.org/10.1128/mSystems.00458-19.

Editor Nicola Segata, University of Trento

Address correspondence to Robert M. Pringle, rpringle@princeton.edu, or Lawrence A. David, lawrence.david@duke.edu.

A.T.R. and T.R.K. contributed equally to this work.

For a commentary on this article, see https://doi.org/10.1128/mSystems.00634-19.

Received 29 July 2019
Accepted 11 September 2019
Published 8 October 2019

Reliable dietary data are needed for human biomedical research and for developing appropriate nutritional recommendations. Methods of diet tracking in both research and clinical contexts frequently depend on self-reporting, whether in the form of diaries in which meals are logged (diet records), prompts to remember foods eaten in the past day (24-h recalls), or surveys that ask individuals to summarize their eating habits over time frames of up to a year (food-frequency questionnaires) (1). However, such human diet assessments have notoriously low accuracy due in part to inaccuracies and bias associated with human memory (2–4). These methods can be so misleading that the majority of diet surveys have been found to routinely misreport caloric intake (2). Furthermore, a greater degree of nutrition education did not improve—indeed, worsened—the accuracy of self-reported diet information (5). Even if diet items are accurately reported, accounts typically lack abundance data (i.e., logs note whether an ingredient was present in the diet but not the amount consumed), and thus, self-reported data are likely to overestimate the importance of rare food items and underestimate common ones. There is therefore a need for alternative methods of quantifying human diet composition (4).

DNA sequencing methods are increasingly used to infer the diets of wild animal populations for which reliable observational data are difficult or impossible to obtain (6). An amplicon-based sequencing technique, known as DNA metabarcoding, is commonly applied in zoology (7–9), microbial community ecology (10), and environmental DNA studies (11) to identify species based on reference databases containing diagnostic sequences (DNA barcodes). Sequencing of plant biomarkers has been used to assess the diet composition of individual herbivore and omnivore species (6, 12–14), to compare diets across species and analyze food web networks (8, 15, 16), and to evaluate differences in food selection by model lab mice under experimentally controlled conditions of nutrient and disease stress (17). Importantly, there is clear potential to apply similar techniques to characterize human diet composition in ways that may support biomedical research and applications (7).

We investigated the utility of DNA metabarcoding for characterizing the plant component of human diets. We applied to human stool samples a widely used protocol for plant DNA metabarcoding, based on amplification and sequencing of the *trn*L(UAA)-P6 marker from chloroplast DNA (6, 11). This marker is useful for dietary analysis due to its short length, conserved primer sites, and interspecific variation (6, 7). It has previously been shown to successfully identify plant DNA in human feces (7) and used to analyze the diet composition of wild herbivores (8, 11, 12, 18). We analyzed samples from a previous diet-intervention study (19) to investigate if (i) self-reported differences in diet composition correspond to DNA-based differences in diet composition and (ii) DNA-based methods can identify experimentally induced dietary changes in diet composition.

RESULTS

We applied DNA metabarcoding to fecal samples from a cohort of 11 individuals who consumed prepared diets with controlled sets of plant ingredients (19). During the study, participants were fed two controlled diets with free eating during a preceding baseline and following washout periods: the plant diet arm included selected grains, legumes, fruits, and vegetables while the animal arm included prepared meats, eggs, and cheeses. We analyzed samples from the end of each diet intervention as well as various free-eating time points (see Fig. S1 in the supplemental material).

In total, we observed a PCR band in 50% of the 54 human samples available from the prepared-diet study. Success varied significantly by diet type ($P = 0.05$, $\chi^2 = 2.83$, DF = 1, chi-square test), with more samples that were collected during the animal diet arm failing to amplify (71%) than those from the plant diet arm (30%). Approximately half of the baseline and washout samples (48%) were successful. From the PCR-positive samples, we obtained 2,113,660 *trn*L-P6 sequence reads that perfectly matched 78 sequences from the reference database. After combining sequences that could not be fully distinguished at the species level (see Materials and Methods), our analyses

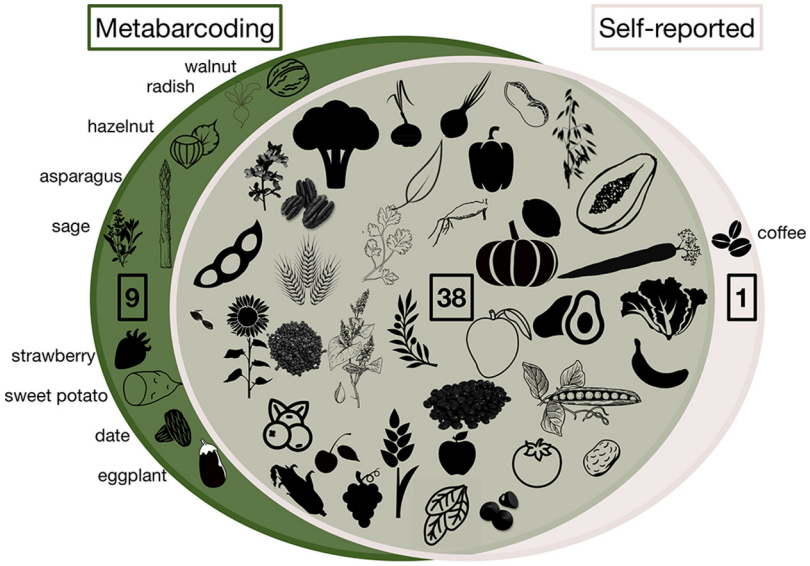

FIG 1 Most plant taxa (79%) were recorded as present at least once in both diet diaries and metabarcoding. Whereas some plants (19%) were found via metabarcoding but not recorded in diaries, only one (coffee) was recorded in diet diaries but absent in metabarcoding. Common names of taxa unique to one method are specified around the Venn diagram.

captured 47 dietary plant taxa. Of these, 39 were identifiable to species level, 4 were identifiable to genus level, and 4 included multiple genera (Table S3). These perfectly matched sequences represented over 70% of the total sequence reads. The median perfect-match read depth was 4,273 per sequence taxon (range = 1 to 556,223).

We compared DNA metabarcoding results to diet diaries kept by participants before, during, and after the controlled-feeding study and found that 38 taxa (79%) appeared in both the sequencing and diary data sets, whereas only one (2%) was solely recorded in the diet diaries (Fig. 1). We next calculated the percentages of plant taxa recorded by participants as having been consumed that were captured by DNA metabarcoding (recall) and the percentage of plant taxa detected by DNA metabarcoding that was also reported in diet diaries (precision). High recall would suggest that metabarcoding yields data that are similar to self-reports. Low precision is harder to interpret, as it could indicate that metabarcoding captured aspects of diet that diaries did not and/or that some proportion of the sequences are false positives. Across all fecal samples, the metabarcoding method had a recall of 0.76 and precision of 0.26 for determining presence/absence of dietary plants in light of the participant's diet record; these two measures are summarized by an F-measure of 0.39 (Table S5). Recall, precision, and F-measure all range from 0 to 1, with 1 representing perfect performance; the F-measure calculated here is unweighted (i.e., assigns equal importance to recall and precision) and is the harmonic mean of recall and precision, which means it tends toward the lesser value of the two. We observed elevated rates of putative false positives for some plants: 25 taxa had false-positive rates greater than 50%.

In fecal samples from the plant-diet arm alone, recall, precision, and F-measure were greater than for the complete data set—0.86, 0.55, and 0.67, respectively (Fig. 2; Table S5). This difference is unsurprising because self-reports are expected to be more accurate during this period of controlled, limited diets and there is also likely higher plant DNA content in stool samples. The only plant-based food present in diet logs that was never detected by metabarcoding was coffee, whereas plants that were inconsistently detected included tea and peppers—in general, beverages and spices may be hard to detect due to low abundance in the diet and high rates of processing.

Coarsening the taxonomic resolution of plant identifications marginally increased the apparent recall of the DNA metabarcoding method (0.73 at species level versus 0.82 at family level) as well as its precision (0.25 versus 0.33, respectively), reflected in an

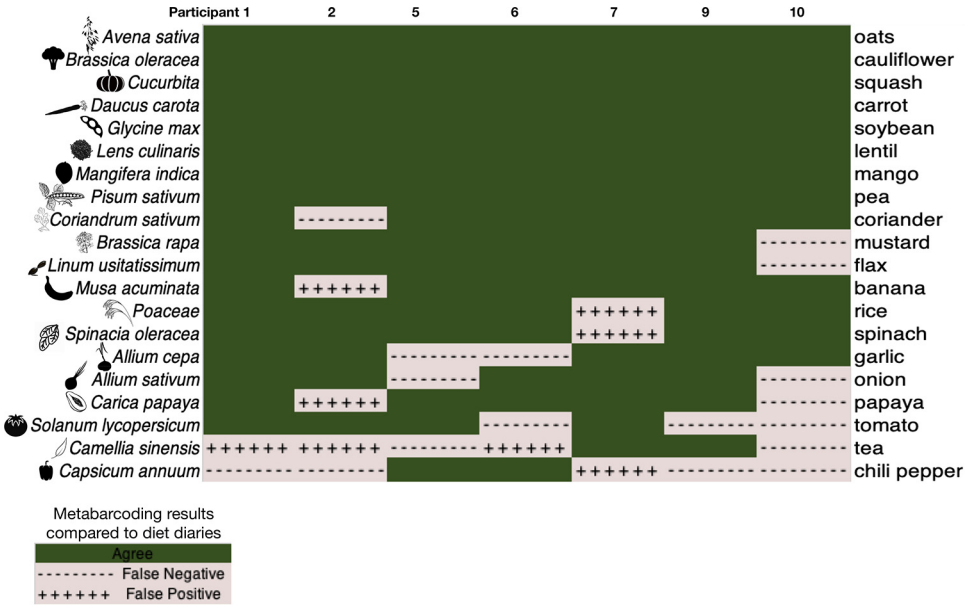

FIG 2 Congruence (green) between diet-diary entries from the day preceding sampling and metabarcoding was common for controlled diet ingredients during the plant-diet arm. Disagreement between metabarcoding data and the dietary diary, either false negative or false positive, is indicated in pink. Latin names of foods are presented to the left of the heat map, and common names are given on the right.

improvement in F-measure (0.37 versus 0.47; Table S5). This was also the case in the plant-diet arm-only samples (recall 0.84 at species level versus 0.92 at family level, precision 0.59 versus 0.59, and F-measure 0.69 versus 0.72). Precision and recall are inversely related; the increase in both metrics that we observed here occurs because the plant taxa involved in the comparison change (in both number and detection status) when they are aggregated to a higher taxonomic level.

We did observe the expected inverse relationship between metrics when the underlying plant taxa remained the same and the detection threshold was varied. Requiring a sequence to exceed a count threshold of 1% or 5% of total reads in order to be defined as present in a given sample led to substantial improvements in precision (increases to 0.51 and 0.51, respectively) but at the cost of recall (decreasing to 0.34 and 0.17, respectively; Table S5). Combined evaluation of these two parameters in the F-measure showed an overall improvement in performance at the 1% threshold (0.41) but a deterioration at 5% (0.25). Interestingly, this trend was not replicated when considering samples from the plant-diet arm only, for which F-measure consistently decreased with an increasing read threshold (to 0.45 at 1% and 0.26 at 5%; derived from recall of 0.30 versus 0.15 and precision of 0.90 versus 0.85 at the 1% and 5% levels, respectively; Table S5). This contrast suggests that imposing a read threshold on the plant-only samples filters out more true positives than false positives and leads to an overall decrease in performance, while a modest read threshold applied to samples including those from nonintervention periods has the opposite effect. This supports the notion that missed reporting of trace plants in diaries but detection by metabarcoding (deemed "false positives" in our analysis framework) has a more prominent effect in freely eaten diets, which included a larger variety of prepared and processed foods that may have obscured these ingredients from the consumer. By comparison, in the plant diet arm, all such diet components were known and could be exhaustively coded from a simply reported menu item (e.g., "Dinner curry") by investigators. Finally, the striking decrease in recall observed in the plant-diet arm samples by applying a 1% read threshold (from 0.86 to 0.30) indicates that true positives are being filtered from the comparison to diet records at this threshold and, thus, that not all low-abundance DNA metabarcoding reads represent false positives.

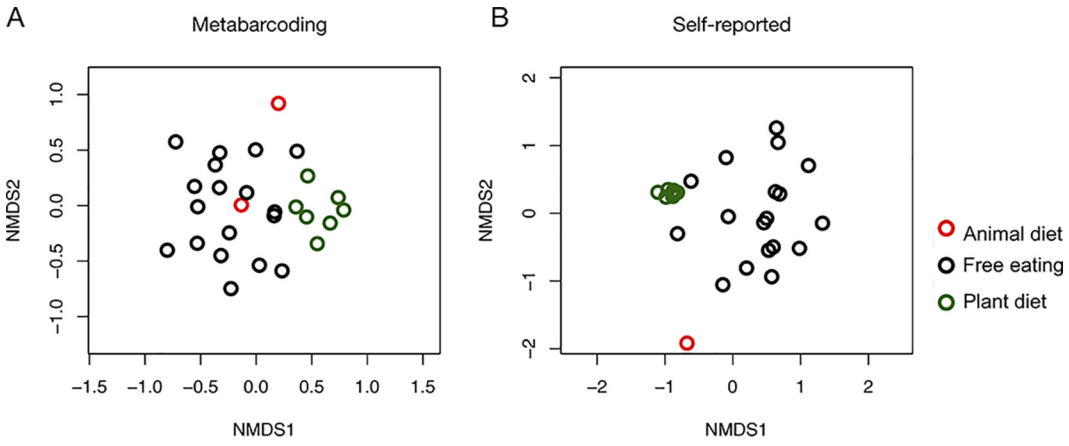

FIG 3 Nonmetric multidimensional scaling (NMDS) of metabarcoding (A) and diet diaries (B) shows separation between experimental diet arms. Samples from participants during the free-eating periods are shown in black ($n = 18$), those from the plant-rich diet period are shown in green ($n = 7$), and those from the animal-rich diet period are shown in red ($n = 2$).

DNA metabarcoding and diary-based methods for characterizing participants' plant intake yielded similar—but nonidentical—results. There was a positive, but weak, correlation between Bray-Curtis dissimilarity of metabarcoding results and data from participant diaries (Mantel statistic = 0.28, $P = 0.002$). We also found that DNA-based dietary composition differed significantly between baseline and experimental diet stages (permutational multivariate analysis of variance [PERMANOVA]: $P < 0.001$, $R^2 = 0.19$, DF = 3, pseudo-F value = 1.80), as visualized using nonmetric multidimensional scaling (NMDS) (Fig. 3). Diary-reported diet composition also differed significantly as a function of experimental diet stage (PERMANOVA: $P < 0.001$, $R^2 = 0.37$, DF = 3, pseudo-F value = 4.58). The two animal-diet samples that we succeeded in amplifying were nearly entirely dissimilar from the plant-diet samples (Bray-Curtis dissimilarity = 0.99 ± 0.01), consistent with the experimental design.

Last, we tested whether differences in plant intake measured by DNA metabarcoding were associated with overall patterns in gut microbial composition or metabolism. We calculated Bray-Curtis dissimilarity matrices based on bacterial relative abundance measured with 16S rRNA gene amplicon sequencing and fecal short-chain fatty acid concentrations (a measure of microbial metabolic functioning) for the baseline samples, during the plant-diet intervention, and during the washout period. Microbial composition was not significantly correlated with either diet self-reports or metabarcoding results at any time point (Mantel tests, $P > 0.05$). Similarly, we did not detect associations between either method of diet analysis and short-chain fatty acid concentrations (Mantel tests, $P > 0.05$). These nonsignificant results may reflect the relative homogeneity of food intake profiles between participants in the plant-diet intervention.

DISCUSSION

We have shown that dietary plant DNA can be amplified and sequenced from human stool using methods commonly applied to wildlife studies. Plant DNA could identify and distinguish experimental and noninterventional diet compositions based on plant taxa commonly consumed by humans. As we were able to detect human consumption of 47 unique plant taxa encompassing 29 plant families, 39 genera, and 39 species, and DNA metabarcoding has previously been employed to characterize the diets of diverse herbivores and omnivores in the wild (8, 13), we believe this approach could be applied effectively to more geographically and culturally disparate human populations in the future.

Before this method is ready for widespread application in biomedical research, further methodological refinements will be necessary. A potentially limited ability to characterize diet composition of free-feeding humans is a challenge that will need to be overcome, because it will often be impossible to distinguish between errors arising

from metabarcoding or diet diaries if the two data sources are in conflict outside experimentally controlled conditions. Both potential sources of error could contribute to the imperfect precision and recall documented here, especially during the free-feeding period of the study. As such, our measurements may best be considered estimates rather than exact precision and recall, as perfect knowledge of participant's diets was unavailable.

Improving the accuracy of human diet diaries may continue to be challenging, owing to the inherent imperfection of memory, but improvements to dietary DNA metabarcoding strategies are occurring rapidly (20, 21). First, many recent improvements to DNA metabarcoding strategies focus on overcoming technical challenges, including optimizing sample handling and extraction, overcoming potential PCR biases, and developing computer algorithms that can more effectively detect and remove aberrant DNA sequences (20, 22, 23). Although our protocol relied on methods that were state of the art at the time, researchers should carefully consider the most recent developments when applying this approach in the future. In particular, further DNA-cleaning protocols to remove polyphenols and other PCR inhibitors commonly found in plants could reduce the rate of PCR failures. Second, researchers are focusing on important considerations related to study design: it is challenging to obtain a highly precise dietary profile from a single sample, and studies pursuing this goal may require a high degree of technical replicates (replicated DNA extractions and PCRs) (24); yet, experimental and computer-simulation analyses suggest that population-level analyses based on well-designed DNA metabarcoding studies can support robust dietary comparisons except in cases of extreme primer bias for or against the most abundant "true" dietary item (20). Despite these potential study limitations, our analysis revealed the expected pattern of nearly complete dietary differentiation between experimental populations that were fed plant- and animal-based diets (Bray-Curtis = 0.99), even with a relatively small sample size ($n = 27$) (Fig. 3).

Important aspects of human physiology and diet composition will be important to consider in the design of DNA metabarcoding experiments that involve people. Diet composition affects gut retention time (25–28), meaning that fecal samples collected simultaneously from two individuals do not necessarily contain foods that the two individuals consumed at the same point in time. DNA copy numbers in fecal samples may also be biased due to differential DNA content in the tissues eaten, digestion of DNA in the gut, and/or recovery of DNA from the resulting specimen. In order to overcome the challenge of discerning how much error exists in the DNA-based analyses and diary-based summaries, future studies should examine large cohorts of people consuming controlled, but varied, diets over time. Although we found DNA from cooked plant material in feces, food preparation and processing could also affect the digestibility of plants (29) and may degrade DNA itself. Notably, coffee—the only plant-based food that was recorded in diaries but never detected by DNA metabarcoding—is derived from seeds that are first roasted and then steeped at high temperatures, all of which could contribute to low quantity and quality of chloroplast DNA markers. Future work should assess how the abundance of DNA markers in feces is impacted by cooking technique and the type of plant tissue consumed. Last, humans consume primarily domesticated plants: only 15 crop species provide almost 70% of the world's calories (30). For example, cruciferous vegetables such as broccoli, kale, and cabbage are all the same species (*Brassica oleracea*) and require extensive sequencing to be distinguished (e.g., 11 to 13 microsatellites) (31, 32); we found here that apples (genus *Malus*) and pears (genus *Pyrus*), as well as rice, rye, and wheat (family Poaceae), are identical at the *trn*L-P6 locus. The use of single-marker loci in DNA metabarcoding studies may therefore be insufficient to differentiate between some foods that are typically considered distinct, including phenotypically and nutritionally variable plants or plant parts, and approaches based on multiple markers warrant exploration. A more diverse reference database would be necessary regardless if this approach were to be applied to human populations who consume more wild plants (33–35).

Despite these current limitations, DNA-based dietary analyses hold promise for

tracking human plant intake. In particular, we believe this approach could be used to increase the frequency with which human plant diet is monitored in biomedical research and clinical applications, as metabarcoding complements standard methods in research on digestion and gastrointestinal health. Fecal samples are regularly collected by medical providers as well as by researchers for microbiome analysis but are to our knowledge not used for dietary sequencing in humans. In the future, DNA metabarcoding could enable investigators to retrospectively infer plant and animal intake among study participants who have banked stool samples but not tracked their diets; such samples are increasingly abundant due to the growing number of human gut microbiome studies (36). Here, the same DNA extractions were used for microbial community profiling and plant metabarcoding. Comparisons between these produced results consistent with the previous finding that the plant-diet experimental treatment was associated with only weak changes in microbiota structure (19). Other applications might include assessing compliance during dietary intervention studies or under restricted diets and overcoming linguistic and other human cultural barriers that prevent accurate communication of diet with self-reporting. Applying DNA metabarcoding to a wider range of human cohorts should be used to determine the utility of the approach for identifying dietary signals diagnostic or causal of various human diseases. Altogether, DNA metabarcoding has become increasingly common in environmental biology (7), and we believe that future applications and refinements of the approach described here could be valuable in studies of human nutrition and health. In conjunction with applying other molecular approaches to human samples, such as microscopy, stable isotope probing, and multi-omics techniques (37–39), a more complete picture of human diets is possible.

MATERIALS AND METHODS

Experimental diet study samples and metadata. Fecal DNA samples were obtained from a previous experimental study on the effects of short-term dietary interventions on the microbiota (19). Analyses were determined to be exempt by the Duke Health Institutional Review Board (Pro00100567). Samples originated from 11 study participants who collected feces each day during 4 days of baseline analysis, 5 days of a plant-based diet, and 6 days of washout and then again for 4 days of baseline, 5 days of an animal-based diet, and 6 days of washout (see Fig. S1 in the supplemental material). The plant-based diet was composed of selected grains, legumes, fruits, and vegetables; the animal-based diet was composed of prepared meats, eggs, and cheeses (Table S1). On both diet arms of the experiment, participants were instructed to eat only study-provided meals and snacks or allowable beverages (water or unsweetened tea for both diets; coffee was allowed on the animal-based diet). They were also allowed to add one salt packet per meal, if desired for taste. Participants could eat unlimited amounts of the provided foods. Participants ate freely during the baseline and washout periods. Across all study days, participants kept daily diet diaries that recorded the quantity and makeup of their unconstrained diets during the baseline/washout periods and, similarly, the quantity and type of the prepared foods they chose to eat during the experimental diet arms. During both free-feeding and experimental diet arms, participants consumed a mix of both cooked and uncooked ingredients, but the preparation method was not always recorded. Rapid and reproducible changes in gut microbiota community structure, gene expression, and metabolism were detected across study participants during diet arms (19), which suggested that participants complied with study diet designs.

Samples were selected for plant DNA metabarcoding from the ends of the baseline period, experimental interventions, and washout periods ($n = 54$ fecal samples; Fig. S1). One participant did not participate in each arm of the experiment, and DNA was no longer available for some participants at certain time points, but we were able to include at least 9 participants from each diet-arm grouping. Diet-diary data were coded from diary entries on the day prior to fecal sample collection. DNA was extracted using a PowerSoil DNA extraction kit (MoBio) and then stored frozen as part of the original study. Data describing gut microbial composition and one measure of microbial function (short-chain fatty acid concentration) were also drawn from the work of David et al. (19). In short, microbial community composition was determined by 16S rRNA gene amplicon sequencing with the Illumina platform. Short-chain fatty acid concentrations were measured with gas chromatography.

DNA metabarcoding sequencing and processing. We used the P6 loop of the chloroplast *trn*L (UAA) intron (*trn*L-P6), which is a broad-spectrum marker useful for DNA metabarcoding of plant species, with published primers (7) and established laboratory protocols (8). Briefly, the *trn*L-P6 locus was amplified with molecular identification (MID) tags to enable pooling and demultiplexing. Pooled amplicons were assembled into a library using the Apollo 324 NGS Library Prep system and PrepX DNA kit (WaferGen, CA), which included DNA end-repairing, A-tailing, adapter ligation, and limited amplification before Illumina barcodes were ligated to the pool for sequencing on an Illumina HiSeq 2500 Rapid Flowcell at Princeton University's Lewis Sigler Institute as single-end 170-nucleotide (nt) reads.

We compiled a reference database comprising the *trn*L-P6 sequences of commonly consumed plant species. To obtain reference sequences, we compiled a list of scientific names from 86 domesticated plant taxa and queried GenBank for records matching "*trn*L" and each of these genus- or species-level groups. A total of 4,688 sequences matching these search terms were downloaded from GenBank in October 2016, and we used the ecoPCR function from the obitools software (40) to search these records for the full-length *trn*L-P6 marker. In this search, we allowed for up to 4 mismatches to the same primers used in metabarcoding analyses and considered sequences spanning 9 to 300 bp in length. We retained reference sequences that were identifiable to genus level using the NCBI taxonomic database. A total of 185 unique sequences representing 2,162 GenBank accessions representing 72 species were obtained from this search for the full-length *trn*L-P6 reference sequence (Data Set S1). The number of sequences in the database exceeds the number of plant species considered in the search because some food species may be represented by multiple haplotypes or because they are represented by congeneric taxa. Based on this database, some common food items are difficult or impossible to distinguish genetically from close relatives despite readily apparent phenotypic differences that can be noted in diet logs (e.g., broccoli, Brussels sprouts, and cabbage [*Brassica oleracea*]; pumpkin and zucchini [*Cucurbita pepo*]; hot and bell peppers [*Capsicum annuum*]; citrus fruits [*Citrus* spp.]); others are phenotypically similar and called the same common name but are different species (e.g., berries that include members of the genera *Rubus*, *Vaccinium*, and *Fragaria* and various species of *Phaseolus* collectively referred to as "beans"). These genetic issues prevented us from identifying some metabarcoding-derived sequences to the species level, and the lexical issues prevented us from identifying some self-reported foods to the species level. Taxa that could not be distinguished by sequence or by name were combined at a higher taxonomic level, and the corresponding entries in diet logs were similarly combined for accurate comparison. These changes affected the taxonomic assignment of 26 unique *trn*L-P6 sequences from the metabarcoding analysis (Tables S2 and S3), and the resulting taxonomic classification was used in all subsequent analyses. In some cases, sequences were unavailable in GenBank or their species-level identifications were deemed uncertain. This affected a few plants found in participant diet logs, including various spices and cranberry, and these taxa were excluded from downstream analyses for both metabarcoding and diet-log analyses (Table S1).

The fecal DNA sequences were demultiplexed and identified through comparison to the reference database. Demultiplexing, identification, and quality controls were performed using obitools software (40). At this stage, we removed sequences with >2 mismatches to the primers, sequences with Illumina fastq quality scores averaging ≤ 32 across the length of the *trn*L-P6 sequence, sequences that contained any ambiguous base calls, and sequences that were <9 bp. We tallied identical sequences in the remaining data set and dropped those that occurred <10 times across all samples that were included in the data set (including controls, extraction blanks, and dietary samples that were subsequently dropped from analysis). A data set of 21,325 unique sequences (2,899,718 total sequence reads) was produced, and only sequences with 100% match identity to a food-plant sequence in the reference database were retained for further analyses ($n = 78$ perfect matches in comparison to the 185 unique *trn*L-P6 sequences in the database).

Analyses. The DNA metabarcoding results were benchmarked for their precision and recall compared to recorded diet. Our benchmarking procedure required assumptions about the completeness of diet records, and because these are known to have frequent inaccuracies (2–4), our results may best be interpreted as estimates of precision and recall. We assumed that omission of foods from diet diaries due to memory lapses, selective reporting, or intake of prepared or processed foods in which not all ingredients were known to the consumer was more likely than the erroneous reporting of a food that was not in fact consumed. Thus, we prioritized metrics that make comparisons between metabarcoding and foods reported as present (rather than absent) in diet diaries. We calculated (i) recall (also called sensitivity), defined as the percentage of foods in diet diaries that were also detected by DNA metabarcoding, and (ii) precision (also called positive predictive value), defined as the percentage of plant taxa detected by DNA metabarcoding that were also recorded in diet diaries. These calculations were performed by comparing diet records that coded a plant taxon as present or absent to the metabarcode read counts that corresponded to the same plant taxon. Because there is an inverse relationship between precision and recall, we also calculated the F-measure, which represents the harmonic mean of precision and recall and ranges from 0 (completely inaccurate detection) to 1 (perfect precision and recall).

For calculation of precision and recall at different taxonomic levels, species were collapsed to shared genera and genera were collapsed to shared families by summing read counts (in the case of metabarcoding data) or by combining binary presence/absence data using an "OR" operator (in the case of reported consumption of a plant taxon in the diet). We repeated this calculation by applying common thresholds of sequence relative read abundance required to infer the "presence" of a plant within a sample (i.e., $>0\%$, 1%, and 5%).

We performed Mantel tests to compare the diets captured by metabarcoding and participant reporting as well as to compare diet summaries and gut microbial composition and functioning. Metabarcoding, microbial composition, and short-chain fatty acid data were processed using the abundance-weighted Bray-Curtis dissimilarity, whereas diet diary data were analyzed only as presence/absence (Jaccard index). Analyses were conducted on each experimental window separately (baseline, plant diet intervention, plant diet washout, and animal diet washout) to exclude multiple measurements of the same individual. Bonferroni corrections were applied to address multiple-hypothesis testing. To determine if metabarcoding and/or participant recording reflected the effect of the experimental diet treatments (free eating, animal diet, or plant diet) we performed permutational multivariate analysis of variance (PERMANOVA). Tests were performed with the vegan package (41) in R (version 3.3) (42).

Data availability. Sequencing data acquired for this study are available through the European Nucleotide Archive under accession number PRJEB34336. The reference sequences are available in Data Set S1.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at https://doi.org/10.1128/mSystems.00458-19.

FIG S1, PDF file, 0.7 MB.
TABLE S1, XLSX file, 0.01 MB.
TABLE S2, XLSX file, 0.02 MB.
TABLE S3, XLSX file, 0.01 MB.
TABLE S4, XLSX file, 0.01 MB.
TABLE S5, XLSX file, 0.01 MB.
DATA SET S1, TXT file, 0.1 MB.

ACKNOWLEDGMENTS

We thank members of the David lab and four reviewers for helpful comments on the manuscript.

L.A.D. acknowledges support from the Triangle Center for Evolutionary Medicine and the Children's Health and Discovery Initiative.

A.T.R., T.R.K., R.M.P., and L.A.D. conceived the project. A.T.R. and T.R.K. carried out molecular experiments. A.T.R., T.R.K., and B.L.P. analyzed sequencing results. P.J.T. provided samples, data, and insight into the design of the diet study. All authors discussed the results and contributed to reviewing and revising the manuscript.

REFERENCES

1. Thompson FE, Subar AF. 2017. Dietary assessment methodology, p 5–48. *In* Coulston A, Boushey C, Ferruzzi M, Delahanty L (ed), Nutrition in the prevention and treatment of disease, 4th ed. Academic Press, Cambridge, MA.
2. Archer E, Hand GA, Blair SN. 2013. Validity of U.S. nutritional surveillance: national health and nutrition examination survey caloric energy intake data, 1971–2010. PLoS One 8:e76632. https://doi.org/10.1371/journal.pone.0076632.
3. Archer E, Pavela G, Lavie CJ. 2015. The inadmissibility of what we eat in America and NHANES dietary data in nutrition and obesity research and the scientific formulation of national dietary guidelines. Mayo Clin Proc 90:911–926. https://doi.org/10.1016/j.mayocp.2015.04.009.
4. Subar AF, Freedman LS, Tooze JA, Kirkpatrick SI, Boushey C, Neuhouser ML, Thompson FE, Potischman N, Guenther PM, Tarasuk V, Reedy J, Krebs-Smith SM. 2015. Addressing current criticism regarding the value of self-report dietary data. J Nutr 145:2639–2645. https://doi.org/10.3945/jn.115.219634.
5. Sugimoto M, Asakura K, Masayasu S, Sasaki S. 2016. Relatively severe misreporting of sodium, potassium, and protein intake among female dietitians compared with nondietitians. Nutr Res 36:818–826. https://doi.org/10.1016/j.nutres.2016.04.011.
6. Pompanon F, Deagle BE, Symondson WOC, Brown DS, Jarman SN, Taberlet P. 2012. Who is eating what: diet assessment using next generation sequencing. Mol Ecol 21:1931–1950. https://doi.org/10.1111/j.1365-294X.2011.05403.x.
7. Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E. 2007. Power and limitations of the chloroplast trnL (UAA) intron for plant DNA barcoding. Nucleic Acids Res 35:e14. https://doi.org/10.1093/nar/gkl938.
8. Kartzinel TR, Chen PA, Coverdale TC, Erickson DL, Kress WJ, Kuzmina ML, Rubenstein DI, Wang W, Pringle RM. 2015. DNA metabarcoding illuminates dietary niche partitioning by African large herbivores. Proc Natl Acad Sci U S A 112:8019–8024. https://doi.org/10.1073/pnas.1503283112.
9. Pringle RM, Kartzinel TR, Palmer TM, Thurman TJ, Fox-Dobbs K, Xu CCY, Hutchinson MC, Coverdale TC, Daskin JH, Evangelista DA, Gotanda KM, Veld N, Wegener JE, Kolbe JJ, Schoener TW, Spiller DA, Losos JB, Barrett R. 2019. Predator-induced collapse of niche structure and species coexistence. Nature 570:58–64. https://doi.org/10.1038/s41586-019-1264-6.
10. Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Lozupone CA, Turnbaugh PJ, Fierer N, Knight R. 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. Proc Natl Acad Sci U S A 108:4516–4522. https://doi.org/10.1073/pnas.1000080107.
11. Taberlet P, Coissac E, Pompanon F, Brochmann C, Willerslev E. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Mol Ecol 21:2045–2050. https://doi.org/10.1111/j.1365-294X.2012.05470.x.
12. Willerslev E, Davison J, Moora M, Zobel M, Coissac E, Edwards ME, Lorenzen ED, Vestergård M, Gussarova G, Haile J, Craine J, Gielly L, Boessenkool S, Epp LS, Pearman PB, Cheddadi R, Murray D, Bråthen KA, Yoccoz N, Binney H, Cruaud C, Wincker P, Goslar T, Alsos IG, Bellemain E, Brysting AK, Elven R, Sønstebø JH, Murton J, Sher A, Rasmussen M, Rønn R, Mourier T, Cooper A, Austin J, Möller P, Froese D, Zazula G, Pompanon F, Rioux D, Niderkorn V, Tikhonov A, Savvinov G, Roberts RG, MacPhee RDE, Gilbert MTP, Kjær KH, Orlando L, Brochmann C, Taberlet P. 2014. Fifty thousand years of Arctic vegetation and megafaunal diet. Nature 506:47–51. https://doi.org/10.1038/nature12921.
13. Soininen EM, Valentini A, Coissac E, Miquel C, Gielly L, Brochmann C, Brysting AK, Sonstebo JH, Ims RA, Yoccoz NG, Taberlet P. 2009. Analysing diet of small herbivores: the efficiency of DNA barcoding coupled with high-throughput pyrosequencing for deciphering the composition of complex plant mixtures. Front Zool 6:16. https://doi.org/10.1186/1742-9994-6-16.
14. Craine JM, Angerer JP, Elmore A, Fierer N. 2016. Continental-scale patterns reveal potential for warming-induced shifts in cattle diet. PLoS One 11:e0161511. https://doi.org/10.1371/journal.pone.0161511.
15. Gebremedhin B, Flagstad O, Bekele A, Chala D, Bakkestuen V, Boessenkool S, Popp M, Gussarova G, Schroder-Nielsen A, Nemomissa S, Brochmann C, Stenseth NC, Epp LS. 2016. DNA metabarcoding reveals diet overlap between the endangered Walia ibex and domestic goats—implications for conservation. PLoS One 11:e0159133. https://doi.org/10.1371/journal.pone.0159133.
16. García-Robledo C, Erickson DL, Staines CL, Erwin TL, Kress WJ. 2013. Tropical plant-herbivore networks: reconstructing species interactions using DNA barcodes. PLoS One 8:e52967. https://doi.org/10.1371/journal.pone.0052967.
17. Budischak SA, Hansen CB, Caudron Q, Garnier R, Kartzinel TR, Pelczer I, Cressler CE, van Leeuwen A, Graham AL. 2018. Feeding immunity:

physiological and behavioral responses to infection and resource limitation. Front Immunol 8:1914. https://doi.org/10.3389/fimmu.2017.01914.
18. Craine JM, Towne EG, Miller M, Fierer N. 2015. Climatic warming and the future of bison as grazers. Sci Rep 5:16738. https://doi.org/10.1038/srep16738.
19. David LA, Maurice CF, Carmody RN, Gootenberg DB, Button JE, Wolfe BE, Ling AV, Devlin AS, Varma Y, Fischbach MA, Biddinger SB, Dutton RJ, Turnbaugh PJ. 2014. Diet rapidly and reproducibly alters the human gut microbiome. Nature 505:559–563. https://doi.org/10.1038/nature12820.
20. Deagle B, Thomas AC, McInnes JC, Clarke LJ, Vesterinen EJ, Clare EL, Kartzinel TR, Eveson JP. 2019. Counting with DNA in metabarcoding studies: how should we convert sequence reads to dietary data? Mol Ecol 28:391–406. https://doi.org/10.1111/mec.14734.
21. Elbrecht V, Vamos EE, Steinke D, Leese F. 2018. Estimating intraspecific genetic diversity from community DNA metabarcoding data. PeerJ 6:e4644. https://doi.org/10.7717/peerj.4644.
22. Schnell IB, Bohmann K, Gilbert M. 2015. Tag jumps illuminated–reducing sequence-to-sample misidentifications in metabarcoding studies. Mol Ecol Resour 15:1289–1303. https://doi.org/10.1111/1755-0998.12402.
23. Coissac E, Riaz T, Puillandre N. 2012. Bioinformatic challenges for DNA metabarcoding of plants and animals. Mol Ecol 21:1834–1847. https://doi.org/10.1111/j.1365-294X.2012.05550.x.
24. De Barba M, Miquel C, Boyer F, Mercier C, Rioux D, Coissac E, Taberlet P. 2014. DNA metabarcoding multiplexing and validation of data accuracy for diet assessment: application to omnivorous diet. Mol Ecol Resour 14:306–323. https://doi.org/10.1111/1755-0998.12188.
25. Wu GD, Chen J, Hoffmann C, Bittinger K, Chen YY, Keilbaugh SA, Bewtra M, Knights D, Walters WA, Knight R, Sinha R, Gilroy E, Gupta K, Baldassano R, Nessel L, Li H, Bushman FD, Lewis JD. 2011. Linking long-term dietary patterns with gut microbial enterotypes. Science 334:105–108. https://doi.org/10.1126/science.1208344.
26. Cummings JH, Jenkins DJ, Wiggins HS. 1976. Measurement of the mean transit time of dietary residue through the human gut. Gut 17:210–218. https://doi.org/10.1136/gut.17.3.210.
27. Cunningham KM, Daly J, Horowitz M, Read NW. 1991. Gastrointestinal adaptation to diets of differing fat composition in human volunteers. Gut 32:483–486. https://doi.org/10.1136/gut.32.5.483.
28. Cummings JH, Hill MJ, Jenkins DJ, Pearson JR, Wiggins HS. 1976. Changes in fecal composition and colonic function due to cereal fiber. Am J Clin Nutr 29:1468–1473. https://doi.org/10.1093/ajcn/29.12.1468.
29. Carmody RN, Wrangham RW. 2009. The energetic significance of cooking. J Hum Evol 57:379–391. https://doi.org/10.1016/j.jhevol.2009.02.011.
30. Ross-Ibarra J, Morrell PL, Gaut BS. 2007. Plant domestication, a unique opportunity to identify the genetic basis of adaptation. Proc Natl Acad Sci U S A 104(Suppl 1):8641–8648. https://doi.org/10.1073/pnas.0700643104.
31. Louarn S, Torp AM, Holme IB, Andersen SB, Jensen BD. 2007. Database derived microsatellite markers (SSRs) for cultivar differentiation in Brassica oleracea. Genet Resour Crop Evol 54:1717–1725. https://doi.org/10.1007/s10722-006-9181-6.
32. Tonguç M, Griffiths PD. 2004. Genetic relationships of Brassica vegetables determined using database derived sequence repeats. Euphytica 137:193–201. https://doi.org/10.1023/B:EUPH.0000041577.84388.43.
33. Şerban P, Wilson JR, Vamosi JC, Richardson DM. 2008. Plant diversity in the human diet: weak phylogenetic signal indicates breadth. Bioscience 58:151–159. https://doi.org/10.1641/B580209.
34. Termote C, Van Damme P, Djailo B. 2011. Eating from the wild: Turumbu, Mbole and Bali traditional knowledge on non-cultivated edible plants, District Tshopo, DRCongo. Genet Resour Crop Evol 58:585–618. https://doi.org/10.1007/s10722-010-9602-4.
35. Schnorr SL, Candela M, Rampelli S, Centanni M, Consolandi C, Basaglia G, Turroni S, Biagi E, Peano C, Severgnini M, Fiori J, Gotti R, De Bellis G, Luiselli D, Brigidi P, Mabulla A, Marlowe F, Henry AG, Crittenden AN. 2014. Gut microbiome of the Hadza hunter-gatherers. Nat Commun 5:3654. https://doi.org/10.1038/ncomms4654.
36. Song SJ, Amir A, Metcalf JL, Amato KR, Xu ZZ, Humphrey G, Knight R. 2016. Preservation methods differ in fecal microbiome stability, affecting suitability for field studies. mSystems 1:e00021-16. https://doi.org/10.1128/mSystems.00021-16.
37. Maixner F, Turaev D, Cazenave-Gassiot A, Janko M, Krause-Kyora B, Hoopmann MR, Kusebauch U, Sartain M, Guerriero G, O'Sullivan N, Teasdale M, Cipollini G, Paladin A, Mattiangeli V, Samadelli M, Tecchiati U, Putzer A, Palazoglu M, Meissen J, Losch S, Rausch P, Baines JF, Kim BJ, An HJ, Gostner P, Egarter-Vigl E, Malfertheiner P, Keller A, Stark RW, Wenk M, Bishop D, Bradley DG, Fiehn O, Engstrand L, Moritz RL, Doble P, Franke A, Nebel A, Oeggl K, Rattei T, Grimm R, Zink A. 2018. The iceman's last meal consisted of fat, wild meat, and cereals. Curr Biol 28:2348–2355. https://doi.org/10.1016/j.cub.2018.05.067.
38. Warinner C, Hendy J, Speller C, Cappellini E, Fischer R, Trachsel C, Arneborg J, Lynnerup N, Craig OE, Swallow DM, Fotakis A, Christensen RJ, Olsen JV, Liebert A, Montalva N, Fiddyment S, Charlton S, Mackie M, Canci A, Bouwman A, Ruhli F, Gilbert MT, Collins MJ. 2014. Direct evidence of milk consumption from ancient human dental calculus. Sci Rep 4:7104. https://doi.org/10.1038/srep07104.
39. Macko SA, Engel MH, Andrusevich V, Lubec G, O'Connell TC, Hedges REM. 1999. Documenting the diet in ancient human populations through stable isotope analysis of hair. Philos Trans R Soc Lond B 354:65–76. https://doi.org/10.1098/rstb.1999.0360.
40. Boyer F, Mercier C, Bonin A, Le Bras Y, Taberlet P, Coissac E. 2016. OBITOOLS: a UNIX-inspired software package for DNA metabarcoding. Mol Ecol Resour 16:176–182. https://doi.org/10.1111/1755-0998.12428.
41. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin PR, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Szoecs E, Wagner H. 2017. vegan: Community Ecology Package, vR package version 2.4-2. https://CRAN.R-project.org/package=vegan.
42. R Core Team. 2017. R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.