The American Journal of Clinical Nutrition 119 (2024) 1301–1308

The American Journal of CLINICAL NUTRITION

journal homepage: https://ajcn.nutrition.org/

Original Research Article

Dietary pattern and diversity analysis using DietDiveR in R: a cross-sectional evaluation in the National Health and Nutrition Examination Survey

Rie Sadohara, David Jacobs, Mark A Pereira, Abigail J Johnson *

Division of Epidemiology & Community Health, University of Minnesota, Minneapolis, MN, United States

A B S T R A C T

Background: There are few resources available for researchers aiming to conduct 24-h dietary record and recall analysis using R.
Objectives: We aimed to develop DietDiveR, which is a toolkit of functions written in R for the analysis of recall or record data collected with the Automated Self-Administered 24-h Dietary Assessment Tool or 2-d 24-h dietary recall data from the National Health and Nutrition Examination Survey (NHANES). The R functions are intended for food and nutrition researchers who are not computational experts.
Methods: DietDiveR provides users with functions to *1*) clean dietary data, *2*) analyze 24-h dietary intakes in relation to other study-specific metadata variables, *3*) visualize percentages of energy intake from macronutrients, *4*) perform principal component analysis or *k*-means clustering to group participants by similar data-driven dietary patterns, *5*) generate foodtrees based on the hierarchical food group information for food items consumed, *6*) perform principal coordinate analysis taking food grouping information into account, and *7*) calculate diversity metrics for overall diet and specific food groups. DietDiveR includes a self-paced tutorial on a website (https://computational-nutrition-lab.github.io/DietDiveR/). As a demonstration, we applied DietDiveR to a demonstration data set and data from NHANES 2015–2016 to derive a dietary diversity measure of nuts, seeds, and legumes consumption.
Results: Adult participants in the NHANES 2015–2016 cycle were grouped depending on the diversity in their mean consumption of nuts, seeds, and legumes. The group with the highest diversity in nuts, seeds, and legumes consumption had 3.8 cm lower waist circumference (95% confidence interval: 1.0, 6.5) than those who did not consume nuts, seeds, and legumes.
Conclusions: DietDiveR enables users to visualize dietary data and conduct data-driven dietary pattern analyses using R to answer research questions regarding diet. As a demonstration of this toolkit, we explored the diversity of nuts, seeds, and legumes consumption to highlight some of the ways DietDiveR can be used for analyses of dietary diversity.

Keywords: ASA24, NHANES, R, diet, hierarchical food classification, foodtree, dietary diversity, food informatics, legumes

Introduction

Data-driven dietary pattern analysis is increasingly being applied to large data sets of 24-h dietary recall data. Efforts such as the proposed Nutrition for Precision Health, powered by the All of Us research program, are expected to collect 24-h dietary recall data that will be made publicly available. Although methods exist to calculate dietary patterns from food frequency questionnaire data [1], using 24-h recall data poses unique challenges. As more and larger data sets are published, there is also a need for reproducibility and the ability to combine data sets. A majority of the published code for the analysis of 24-h dietary data is written for proprietary software, posing issues for open data science practices. For example, code to process Automated Self-Administered 24-h (ASA24) Dietary Assessment Tool is available only for SAS (SAS Institute Inc.). R is an open-source platform with countless packages and public resources that can be tailored to one's statistical analysis and visualization needs [2]. Therefore, using R for dietary data analysis will increase accessibility and reproducibility, as R is available without a license fee. Despite the benefits to researchers gained by using R, resources are scarce for analyzing 24-h recall data using R; therefore, we present a dietary analysis toolkit, DietDiveR, currently supporting ASA24 and NHANES data. DietDiveR contains R functions for data loading and preparation, exploratory data overview, clustering analysis, foodtree generation, ordination, and diversity metrics (Table 1) with an

Abbreviations: ANCOVA, analysis of covariance; ANOVA, analysis of variance; ASA24, Automated Self-Administered 24-h Dietary Assessment Tool; IPR, income-to-poverty ratio; PCoA, principal coordinate analysis; VVKAJ, Vegetarian; Vegan, Keto; American, and Japanese.

* Corresponding author.
E-mail address: abbyj@umn.edu (A.J. Johnson).

https://doi.org/10.1016/j.ajcnut.2024.02.014
Received 8 August 2023; Received in revised form 8 February 2024; Accepted 16 February 2024; Available online 11 April 2024

TABLE 1
DietDiveR features

Functionality	Description
Load raw data	Load ASA24 data (.csv) or NHANES (.XPT) into R, compute intake day totals from items-level data, and compute means across record collection days
Clean data	Remove incomplete data, apply filters to participants and to dietary data according to specified conditions and/or ASA24 data cleaning guidelines
Summarize data	Generate basic statistics of variables from daily totals data, and generate boxplots and scatterplots of variables of interest
Generate variables to group participants	Generate categorical variables based on demographic, health, or other metadata used to group participants
Visualize percent energy from macronutrients	Compute percent of energy intake from carbohydrate, protein, and total fat, and visualize macronutrient proportions with barplots
Prepare for clustering	Keep complete rows and remove variables with zero variance and optionally remove highly correlated variables
Perform clustering analyses	Perform PCA or *k*-means analyses with nutrient or food group data and generate biplots and barplots of variables contributing to PCs
Generate foodtrees	Generate foodtrees of consumed food items that are grouped according to their similarity to one another
Visualize foodtrees	Visualize foodtrees in a circular plot at a desired classification depth
Perform ordination	Compute distances between participants and perform clustering analyses such as PCoA with the hierarchical information from food groups taken into consideration. Test differences between participant groups by PERMANOVA
Visualize ordination results	Generate biplots with participants color-coded by a participant grouping variables of interest
Inspect food items correlated with ordination axes	Show correlation coefficients and *q*-values for food items and selected ordination axes
Compute diversity of all or specific food groups	Compute alpha-diversity of food items consumed

Abbreviations: ASA24, Automated Self-Administered 24-h dietary assessment tool; PC, principal component; PCA, principal component analysis; PCoA, principal coordinate analysis.

aim to facilitate data-driven dietary pattern analysis by beginner–intermediate R users. We applied DietDiveR to 24-h dietary records data to demonstrate the basic functions of the toolkit. We also applied DietDiveR to a subset of 2-d 24-h recall data from NHANES to demonstrate analyses with dietary diversity metrics. We present the results of this NHANES analysis comparing legume diversity with waist circumference.

Methods

Demonstration data set

Data creation

DietDiveR provides example scripts to prepare, visualize, and analyze dietary data as dietary patterns and dietary diversity (Figure 1). For ASA24 data analysis demonstrations, we created a data set of dietary records (referred to as VVKAJ). VVKAJ is a hypothetical study with 3-d of dietary records from 17 fictitious participants following 1 of 5 diets: Vegetarian, Vegan, Keto, American, and Japanese. Menus for these demonstration data were selected from online blogs and recipes, with a goal of creating meal patterns that would differ by macronutrient profile and food selection (see **Supplementary Methods** for more details).

Data preparation

Using functions from DietDiveR we loaded data and filtered outliers and missing data, computed totals for food groups, nutrients, and other food components for each participant starting from food-level data for each intake day. We similarly computed the average of food groups, nutrients, and other food components across intake days for each individual where multiple days of intake data were available, and performed quality control procedures. Although DietDiveR does not make specific recommendations for quality control, quality control of mean total intake of food group or nutrient data can be completed according to the guidance such as that provided by ASA24 [3], or by using researcher-defined thresholds for nutritional variables of interest.

Clustering analysis

We used DietDiveR functions and filtered variables with zero variance and removed highly correlated variables with a correlation coefficient threshold of $r > 0.75$. We used the set of remaining variables with no missing data (that is, no not available or NA values) and nonzero variance as an input for *k*-means and principal component analyses. DietDiveR uses functions from the ggplot2 [4] and factoextra [5] packages to visualize those clustering results in the form of biplots, scatter plots, and bar plots, with the participants color-coded by their groups.

Tree-based analysis

DietDiveR includes functions to generate foodtrees, where food items are hierarchically grouped [6], such that 9 categories of foods branch into more detailed classifications. By taking into account the food items' phenetic relationships, the participant's dietary patterns can be analyzed not only by the food items they consumed, but also by the similarity or food group of the foods they consumed. Foodtree visualizations can be generated with color-coded food groups to visually explore the participant's overall intake by using ggtree [7]. We used phyloseq and vegan [8,9] to perform foodtree-based principal coordinate analysis (PCoA) with weighted or unweighted UniFrac beta-diversity distances [10] between participants. After checking for differences in beta dispersion (multivariate homogenity of group dispersion), we tested for between-group dietary separation using a permutation-based statistical test (PERMANOVA, 5000 permutations) and the pairwiseAdonis function [11].

NHANES data set

Record filtering

Food item records were obtained from the NHANES 2015−16 24-h recall data. Population-wide adjustment using survey weights was omitted in this example for simplicity of the demonstration; however, researchers should consult NHANES documentation to appropriately incorporate sample weights into their analysis. Food groups, nutrients,

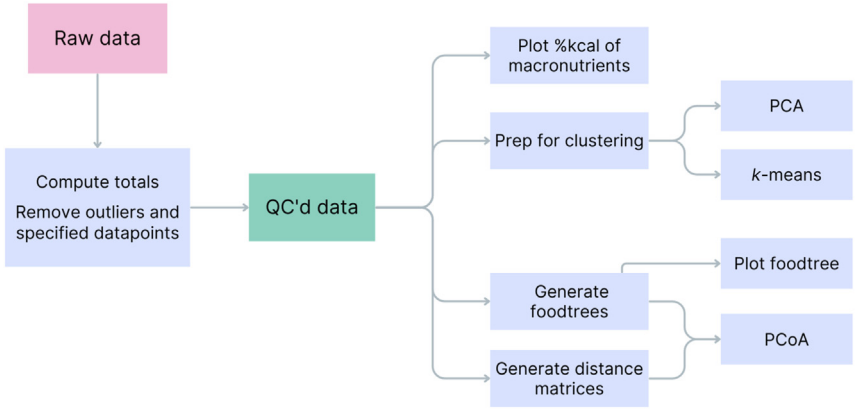

FIGURE 1. Flowchart of dietary data analyses using DietDiveR. %kcal, percentage of in energy intake (kcal); PCA, principal component analysis; PcoA, principal coordinate analysis; QC, quality control.

and other food components data were summed, and the average of the 2-d was calculated. The data set was processed to remove participants with incomplete or potentially inaccurate data for analysis. A detailed flowchart of participant selection is shown in Supplemental Figure 1 and discussed in more detail in results. To remove potential inaccurate dietary records, the 2-d means of totals were cleaned in accordance with the General Guidelines for Reviewing & Cleaning Data by ASA24 [3] using cut points based on the 5th and 95th percentile from NHANES data to identify potential outliers by energy, protein, total fat, and vitamin C intake. Males were removed if their record(s) contained total energy <650 kcal or >5700 kcal, total protein <25 g or >240 g, total fat <25 g or >230 g, and vitamin C <5 mg or >400 mg. Females were removed if their record(s) contained total energy <600 kcal or >4400 kcal, total protein <10 g or >180 g, total fat <15 g or >185 g, or vitamin C <5 mg or >350 mg. Criterion for β-carotene (<15 μg or >8200 μg for males and <15 μg or >7100 μg for females) was not used. Participants with complete BMI and waist circumference data were retained. The sociodemographic variables, age, sex (reported as male, female, or missing), ethnicity (reported as Mexican American, Other Hispanic, Non-Hispanic White, Non-Hispanic Black, Non-Hispanic Asian, Other, or Missing), income (as ratio of family income to poverty), and education (reported as less than 9th grade, 9–11th grade, high school graduate or general educational development (GED) equivalent, some college or AA degree, college graduate or above, refused, do not know, or missing) were included in the NHANES 2015-16 data and were used as covariates.

Grouping by the diversity in nuts,seeds,and legumes consumption

Food items from NHANES with Food and Nutrient Database for Dietary Studies (FNDDS) foodcodes between 40000000 and 49999999, which are nuts, seeds, and legumes (refered to as nuts/seeds/legumes throughout methods and results), were selected from the quality controlled food items of the data set. An individual food consumption table was generated with nuts/seeds/legumes consumed from the 2-d for each participant, and the average consumption (g/d) was computed.

Sociodemographic and nutrients variables

The participants were grouped into 3 age groups: 18–39, 40–59, and 60-y or older; 3 income groups: Family income-to-poverty ratio (IPR) < 1.85, 1.85–2.99, and ≥3.00; and 3 education groups: <high school, high school graduate or some college, and college graduate or above (Table 2). The means of the quantity of protein foods excluding legumes, legumes, total protein foods, the percentages of legumes (if consumed) in total protein food consumption, total vegetable intake except legumes, and total fruit intake were computed for each of the diversity groups. Macronutrients (carbohydrate, protein, and total fat), dietary fiber, total saturated fat, total unsaturated fat (mono- and polyunsaturated fats), alcohol, and added sugar intakes were adjusted per 2000 kcal (Table 3).

Analysis of covariance for body measurements

To analyze body measurements (BMI and waist circumference) for each diversity group, the car package [12] in R was used to calculate type III analysis of covariance (ANCOVA) with diversity group as the treatment effect and age, sex, ethnicity, income, education, and kcal intake as covariates. The estimated marginal means (emmeans) were used to statistically test for associations between waist circumference or BMI and the diversity groups and conduct pairwise comparisons between diversity groups by using the emmeans package in R [13].

Human subjects

The demonstration data set of VVKAJ wascreated for demonstration purposes; it does not contain any real data collected from human subjects. The NHANES 2015–2016 data are publicly available and were downloaded and used in a deidentified form, and no NHANES restricted variables were used; thus, no institutional review was completed.

Software

R version 4.1.2 [2] and Rstudio 2021.09.1+372 "Ghost Orchid" Release" [14] were used to develop DietDiveR and for all statistical analysis. The dependency R packages for DietDiveR are listed in Supplemental Table 1 and **Sample Code** is included in the Supplementary Materials.

Results

Demonstration data highlights the utility of DietDiveR

Exploration of the demonstration data set with DietDiveR showed expected variability in macronutrient content (Figure 2A) and separation in principal component space (Figure 2B). A foodtree was created from the data (Figure 2C), and this was used to show separation in a foodtree-based PCoA ($P = 2 \times 10^{-4}$, PERMANOVA, $\alpha = 0.05$) (Figure 2D).

TABLE 2
Sociodemographic features of the diversity groups

	DivNo (n = 1819)	Div0 (n = 1105)	Div1 (n = 360)	Div2 (n = 357)
	%			
Sex				
Male	49	46	46	45
Female	51	54	54	55
Age, y				
18–39	39	34	32	32
40–59	31	33	37	35
60+	30	33	31	34
Ethnicity				
Mexican American and Hispanic	27	35	28	30
NH White	37	36	38	36
NH Black	26	17	13	9
NH Asian	6	8	19	23
Other	4	4	3	2
Family IPR				
<1.85	49	44	33	32
1.85–2.99	21	20	22	18
≥3.00	29	35	45	50
Education				
<HS	21	21	13	13
HS grad or some collage	59	52	46	40
College grad or above	20	28	41	47

Abbreviations: Div0, 1 nuts/seeds/legumes (no diversity); Div1, diversity of nuts/seeds/legumes below the median; Div2, diversity of nuts/seeds/legumes above the median; DivNo, no nuts/seeds/legumes; HS, high school; IPR, family-wise income-to-poverty ratio; NH, non-Hispanic.

Description of the NHANES data set

NHANES 2015–2016 24-h recall data from day 1 was available for 8505 participants and from day 2 for 7027 participants (with a subset of participants having data for both days). For day 1, 179 participants had incomplete data, 2 reported only 1 food item/d, and 3058 were under 18-y old and were removed, resulting in 5266 day 1 recalls. For day 2, 152 participants had incomplete data, 6 reported only 1 food item/d, and 2468 were under 18-y old and were removed, resulting in 4401 day 2 recalls. The 4401 participants with 2-d of intake data were retained, and the 865 that had only day 1 records were removed. There were a total of 4164 participants (2096 males and 2305 females) with 2-d of dietary data. After an assessment of potential outliers using nutrient value cut points, 92% of male and 95% of female participants with dietary data were retained. After the removal of participants with

TABLE 3
Dietary features of the diversity groups, consumption per day, means, and (SD)

Variable	Unit	DivNo (n = 1819)	Div0 (n = 1105)	Div1 (n = 360)	Div2 (n = 357)	P ANOVA[1]
Food intake						
Legume intake	oz.	0.2 (0.6)	0.7 (1.1)	1.1 (1.5)	1.4 (1.9)	<0.0001
PF + legumes	oz.	5.8 (3.4)	7.0 (3.4)	8.2 (3.8)	8.7 (4.3)	<0.0001
Legumes in total PF intake	%	—	11.1 (16.2)	15 (18.8)	17.4 (20.8)	<0.0001[2]
Total fruit	cup	0.8 (0.9)	1.0 (1.0)	1.2 (1.2)	1.4 (1.4)	<0.0001
Total vegetables excluding legumes	cup	1.4 (0.9)	1.5 (1.0)	1.8 (1.0)	1.7 (1.0)	<0.0001
Nuts/seeds/legumes consumption	g/d	0 (0)	49 (57)	104 (102)	131 (107)	<0.0001[2]
Total number of items reported	ct.	14 (5)	16 (5)	18 (5)	19 (5)	<0.0001
Total energy	kcal	1960 (714)	2031 (694)	2136 (691)	2192 (731)	<0.0001
Nutrient intake[3]						
Carbohydrate	g	241 (45)	238 (42)	239 (45)	239 (45)	0.476
Protein	g	79 (22)	82 (21)	82 (20)	82 (19)	<0.001
Dietary fiber	g	15 (6)	18 (7)	22 (9)	24 (9)	<0.0001
Total fat	g	77 (16)	78 (16)	78 (17)	80 (18)	<0.01
Total saturated fat	g	26 (7)	25 (7)	23 (7)	23 (7)	<0.0001
Total monounsaturated fat	g	27 (6)	28 (7)	28 (8)	30 (9)	<0.0001
Total polyunsaturated fat	g	17 (6)	18 (6)	20 (7)	20 (6)	<0.0001
Total unsaturated fat	g	44 (11)	46 (11)	48 (12)	50 (14)	<0.0001
Alcohol	g	7 (17)	6 (14)	6 (14)	5 (12)	0.095
Added sugars	tsp.	17 (11)	14 (8)	12 (8)	11 (7)	<0.0001

Abbreviations: ANOVA, analysis of variance; Div0, 1 nuts/seeds/legumes (no diversity); Div1, diversity of nuts/seeds/legumes below the median; Div2, diversity of nuts/seeds/legumes above the median; DivNo, no nuts/seeds/legumes; PF, protein food.

[1] P value for ANOVA test for diversity groups.

[2] Tested with Div0, Div1, and Div2, excluding DivNo.

[3] Adjusted per 2000 kcal.

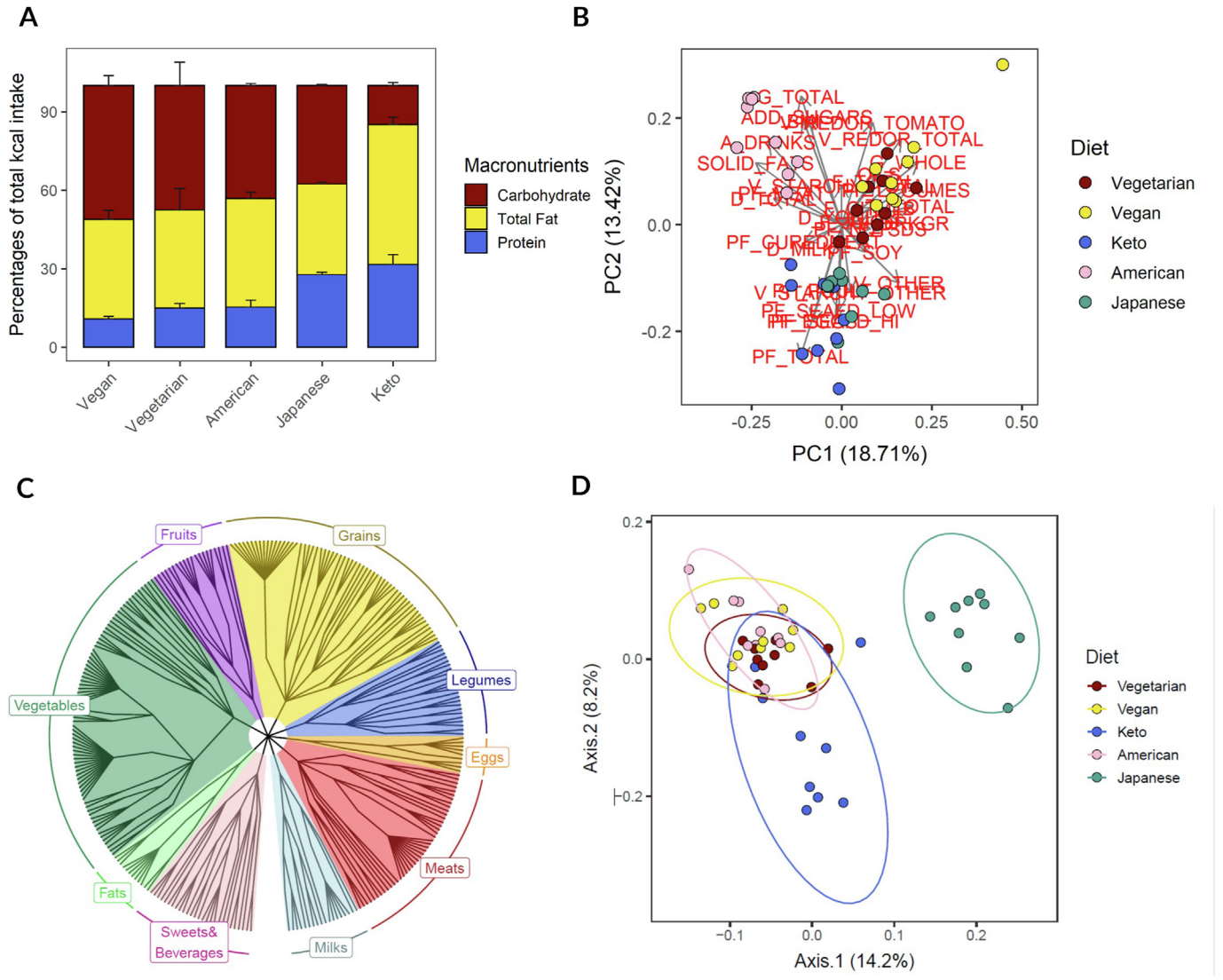

FIGURE 2. Visualizations that can be generated by analyzing the Vegetarian, Vegan, Keto, American, and Japanese diets (VVKAJ) data set using the functions in DietDiveR. (A) Percentages of total calorie intake by macronutrients for various diet groups. (B) Biplot of principal component analysis (PCA) carried out with the food consumption data by food category of the participants. (C) Foodtree of the reported foods by the participants organized by the first level of food categories: milks, meats, eggs, legumes, grains, fruits, vegetables, fats, and sweets, and beverages. (D) Biplot of principal coordinate analysis (PCoA) carried out with weighted UniFrac distance of reported food items and their consumption amounts by the participants. At least 1 diet group is different from the rest (PERMANOVA, $P = 2 \times 10^{-4}$).

missing data for BMI, waist circumference, and income, 3642 participants remained. The ratio of family income-to-poverty (Family IPR) was used as the income variable, and 356 records with missing data for this variable were excluded. The other demographic variables did not have missing data (Supplemental Figure 1). The average nuts/seeds/legumes intake were manually inspected for all remaining participants, and 1 individual with an implausible intake amount (1500 g/d) of nuts/seeds/legumes consumption was considered to be an outlier and was excluded. All other records had nuts/seeds/legumes consumption <1000 g/d. As a result, 3641 records were included in analysis.

Diversity of specific food groups can be calculated from NHANES data

Out of the 3641 participants, 50% ($n = 1822$) reported consuming ≥1 nuts/seeds/legumes food item during the 2-d of NHANES records. They reported consuming a total of 3460 nuts/seeds/legumes food items, and there were 237 unique foods. Alpha diversity using the Shannon index for nuts/seeds/legumes was calculated for each individual. Shannon diversity ranged from 0 to 1.95 (Table 4), and the data were divided into groups for analysis based on this index. Those who

TABLE 4
Shannon's diversity of the diversity groups

Diversity group	*n*	Nuts/seeds/legumes food item consumed	Shannon's diversity index
DivNo	1819	0	NA
Div0	1105	1	0
Div1	360	>1	0.027–0.659
Div2	357	>1	0.660–1.95

Div0, 1 nuts/seeds/legumes (no diversity); Div 1, diversity of nuts/seeds/legumes below the median; Div 2, diversity of nuts/seeds/legumes above the median; DivNo, no nuts/seeds/legumes; NA, not applicable.

consumed no nuts/seeds/legumes foods in the 2-d ($n = 1819$) were grouped as DivNo. Those who consumed only 1 nuts/seeds/legumes food ($n = 1105$) had a diversity index of 0 (only 1 food, thus no diversity), and they were labeled as Div0. Those who consumed >1 nuts/seeds/legumes food ($n = 717$) with a diversity index >0 were split into the lower and upper halves, which were named Div1 and Div2, respectively (Table 4, Supplemental Figure 1). Div1's Shannon diversity ranged from 0.027 to 0.659, and Div2's Shannon diversity ranged from 0.660 to 1.95 (Table 4).

The sociodemographic features such as age, ethnicity, income, and education of the participants are presented in Table 2. The proportion of Asians was higher among the groups with more diversity in nuts/seeds/legumes consumption (6% in the DivNo group and 23% in the Div2 group). Notably, 29% of DivNo participants, who did not consume nuts/seeds/legumes, had Family IPR ≥ 3.00, whereas 50% of Div2 group participants had Family IPR ≥ 3.00. For education, 20% of the DivNo participants had a college degree or higher, whereas 47% of Div2 group participants did. Therefore, high family income and education were associated with higher consumption of and diversity in nuts/seeds/legumes.

Food group diversity and dietary nutrient composition

Diets of the diversity index groups differed by many food groups and nutrients (Table 3). The diversity groups with highest consumption of nuts/seeds/legumes ate more legumes and had higher percentages of legumes in protein foods than the low diversity groups, as expected. Interestingly, the higher the nuts/seeds/legumes diversity groups consumed more vegetables and fruits and total energy [DivNo: 0.8, Div0: 1.0, Div1: 1.2, Div2: 1.4 cups for total fruit consumption; DivNo: 1.4, Div0: 1.5, Div1: 1.8, Div2: 1.7 cups for total vegetables excluding legumes; and DivNo: 1960, Div0: 2031, Div1: 2136, Div2: 2192 kcal for total energy; *P* Analysis of Variance (ANOVA) <0.0001 for all 3]. Because of these differences, nutrient and food component intake was adjusted per 2000 kcal. After adjustment, carbohydrate intake was not different among the diversity groups (DivNo: 241, Div0: 238, Div1: 239, Div2: 239 g; *P*-ANOVA = 0.476), but protein, dietary fiber, total fat, saturated fat, total unsaturated fat, and added sugar intakes were different such that the more diverse nuts/seeds/legumes consumption coincided with more dietary fiber, total fat, total unsaturated fat, and less total saturated fat and added sugars consumption. Interestingly, alcohol consumption was similar among the diversity groups (DivNo: 7, Div0: 6, Div1: 6, Div2: 5 g; *P*-ANOVA = 0.095).

Legume diversity is associated with a lower waist circumference

We modeled the association between waist circumference and nuts/seeds/legumes diversity, incorporating covariates (age, sex, ethnicity, income, education, and energy intake). ANCOVA indicated all the terms were significantly associated with waist circumference, except income and energy ($P = 0.238$ and $P = 0.102$, respectively; Table 5). Waist circumferences were different among the diversity groups ($P < 0.01$, Table 5), and the group that did not consume nuts/seeds/legumes had 3.8 cm larger waist circumference than those who consumed the most diverse nuts/seeds/legumes foods ($P < 0.001$, Tukey-adjusted, Table 6). In addition, the Div0 group that consumed only 1 nuts/seeds/legumes food had a 3.4 cm larger waist circumference than Div2 ($P < 0.01$, Tukey-adjusted, Table 6). Similar results were found for BMI (see Supplementary Results, Supplemental Tables 2 and 3).

TABLE 5
ANCOVA table for waist circumference

	Sum Sq	Df	*F* value	*P* value
(Intercept)	968,293	1	3684	<0.0001
Diversity group	3963	3	5	<0.01
Age, y	37,101	2	71	<0.0001
Sex	5027	1	19	<0.0001
Ethnicity	31,716	4	30	<0.0001
Family IPR	755	2	1	0.238
Education	4204	2	8	<0.001
Kcal	701	1	3	0.102
Residual	952,704	3625		

Abbreviations: ANCOVA, analysis of covariance; Df, degrees of freedom; Kcal, kilocalorie; IPR, income-to-poverty ratio.

TABLE 6
Estimated marginal means (emmean) and contrast of waist circumference (cm) of the 4 levels of the diversity groups

Diversity group	Contrast	emmean	95% CI	SE
DivNo	—	99.8	(98.4, 101.2)	0.5
Div0	—	99.5	(97.8, 101.1)	0.6
Div1	—	99.3	(96.7, 101.8)	0.9
Div2	—	96.1	(93.4, 98.7)	0.9
—	DivNo–Div0	0.4	(−1.4, 2.1)	0.6
—	DivNo–Div1	0.5	(−2.2, 3.2)	1.0
—	DivNo–Div2	3.8	(1.0, 6.5)	1.0
—	Div0–Div1	0.2	(−2.6, 3.0)	1.0
—	Div0–Div2	3.4	(0.6, 6.2)	1.0
—	Div1–Div2	3.2	(−0.2, 6.6)	1.2

Abbreviations: CI, confidence interval; Div0, 1 nuts/seeds/legumes (no diversity); Div1, diversity of nuts/seeds/legumes below the median; Div2, diversity of nuts/seeds/legumes above the median; DivNo, no nuts/seeds/legumes.

Discussion

The visualized proportion of energy from macronutrients and the separation of participants by dietary pattern using PCoA demonstrates the utility of DietDiveR to analyze dietary data collected wtih ASA24 or from NHANES. The DietDiveR tookit provids researchers with tools for general summarization, clustering analyses, incorporation of hierarchical grouping of food items, and ordination analyses using food grouping information. Furthermore, DietDiveR provides researchers with a tool to calculate dietary diversity for overall diet and within specific food groups or categories.

The importance of diet diversity for adequate nutrition intake has long been discussed in the literature, and dietary diversity indices have been used to assess micronutrient adequacy, especially in developing countries [15]. However, the relationship between dietary diversity indices and health outcomes such as noncommunicable diseases is not straightforward [16], with the potential for high dietary diversity to correlate with high energy intake. Alpha-diversity metrics may have important limitations when applied to diet. For example, a diet high in a diverse selection of breads, doughnuts, pastries, and candies, plus sugar sweetened beverages is unlikely to be healthy. It could be possible to see high diversity in both low- and high-quality dietary patterns. Using diversity of different food groups somewhat alleviates these concerns. As this example highlights, these metrics can be hard to describe and understand, which underscores the need for reproducible, published methods with available code to ensure that nuance can be explored.

Here, we preliminarily examined nuts, seeds, and legumes diversity and demonstrated it as a potentially useful index to explore the health-promoting effects of nuts, seeds, andlegumes consumption [17].

We defined 4 dietary intake groups based on legume alpha-diversity (Shannon's diversity, intake of multiple different nuts, seeds, and legumes consumed over 2-d) for $n = 3641$. The highest legume α-diversity group, Div2, had a lower waist circumference and a lower BMI, after adjustment for age, sex, ethnicity, and education variables. This association encourages more studies to be conducted to examine the hypothesis that the consumption of diverse nuts,seeds, andlegumes foods may be associated with a lower risk of developing cardio-metabolic diseases, as a reduction in waist circumference is important to the prevention of cardio-metabolic diseases [18]. In addition, the Div2 group had healthier dietary patterns that are high in legumes, total fruit, and vegetables, the consumption of which are encouraged by Dietary Guidelines for Americans [19]. The Div2 group also had less saturated fat and added sugar, and higher unsaturated fat intake, suggesting that the diversity in nuts,seeds, andlegumes consumption was corrrelated with other healthy dietary features.

Taken together, the findings support the DietDiveR approach, which is to apply ecological metrics to dietary data to calculate dietary alpha- and beta-diversity and to generate diversity of food groups. DietDiveR provides a set of streamlined R functions to clean, analyze, and visualize dietary data from ASA24 and NHANES. By performing all the procedures in R, users can produce consistent and customized outputs with the flexibility that R, an open-source tool with global users, offers.

DietDiveR scripts are highly customizable and require close attention from researchers. Although we provide examples here for data quality control and cleaning, it is noteworthy that in our analysis of NHANES data, 1 outlier with implausible intake of nuts, seeds, andlegumes foods was not initially identified using macro- and micronutrient cut points. This underscores the need for researchers to ensure that their analysis pipelines are well defined and specific to their research question.

We caution any causal or etiological interpretation of the data presented here as we did not incorporate physical activity, or other habits such as smoking or drinking as covariates. Our goal is to demonstrate the capabilities of DietDiveR with an example analysis. We recognize that the results are presented without full incorporation of potential confounders. Energy intake may be underreported in NHANES and the level of underreporting may be correlated with weight status. Exercise levels could be related to waist circumference and BMI as well as diversity in nuts, seeds, and legumes consumption, as those who have higher healthy diet scores tend to be health-conscious and to follow regular exercise routines [20,21].

DietDiveR is an open-source analysis toolkit that offers functions to analyze dietary data collected with ASA24 or from NHANES including data cleaning, clustering analyses, and foodtree-based analyses. Moreover, DietDiveR provides users with tools to calculate data-driven dietary patterns and diversity metrics of specific food groups. Although DietDiveR is not yet available in CRAN or Bioconductor, the toolkit of analysis scripts is accessible to researchers on GitHub. Our example analyses presented here highlight that data-driven dietary patterns can be calculated from food-level data and that ecological diversity metrics applied to dietary data can reveal interesting relationships between food and health and may be useful when applied to future studies.

Acknowledgments

We would like to thank Mo Hutti for the create_corr_frame function that generates a correlation table with ordination axes and variables; Pajau Vangay for the collapse_by_correlation function that removes correlated variables; and Suzie Hoops for matrix multiplication implementation and insights into statistical analyses.

Author contributions

The authors' contributions were as follows – RS, AJJ: conceptualized and created DietDiveR and wrote the manuscript; RS: analyzed the data; RS, AJJ: had primary responsibility for final content; MAP, DJ: were involved in the development of statistical models and data interpretation, and critically revised the manuscript; and all authors: read and approved the final manuscript.

Conflict of interest

MAP is an Editor on *The American Journal of Clinical Nutrition* and played no role in this Journal's evaluation. The other authors report no conflicts of interest.

Funding

This work was supported by start-up funds provided to Abigail Johnson from the University of Minnesota.

Data availability

DietDiveR, including the ASA24 demonstration data, is available for download at https://github.com/computational-nutrition-lab/DietDiveR. The tutorial can be accessed on https://computational-nutrition-lab.github.io/DietDiveR/. The National Health And Nutrition Survey data can be accessed at https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?BeginYear=2015.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.ajcnut.2024.02.014.

References

[1] D.B. Panagiotakos, C. Pitsavos, Y. Skoumas, C. Stefanadis, The association between food patterns and the metabolic syndrome using principal components analysis: the ATTICA Study, J. Am. Diet. Assoc. 107 (6) (2007) 979–987.
[2] R Core Team, R: a language and environment for statistical computing [Internet] R Foundation for Statistical Computing, Vienna, Austria, 2021 [cited February 8, 2023]. Available from: https://www.r-project.org/.
[3] CDC National Center for Health Statistics, Reviewing & cleaning ASA24 data [Internet], General Guidelines for Reviewing & Cleaning Data, 2020 [cited April 5, 2022]. pp. 1–6. Available from: https://epi.grants.cancer.gov/asa24/resources/asa24-data-cleaning-2020.pdf.
[4] H. Wickham, ggplot2: elegant graphics for data analysis [Internet], Springer-Verlag New York, 2016 [cited February 8, 2023]. Available from: https://ggplot2.tidyverse.org.
[5] A. Kassambara, F. Mundt, factoextra: extract and visualize the results of multivariate data analyses [Internet], 2020 [cited February 8, 2023]. Available from: https://cran.r-project.org/package=factoextra.
[6] A.J. Johnson, P. Vangay, G.A. Al-Ghalith, B.M. Hillmann, T.L. Ward, R.R. Shields-Cutler, et al., Daily sampling reveals personalized diet-microbiome associations in humans, Cell, Host Microbe 25 (6) (2019) 789–802.e5.

[7] 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 (1) (2017) 28–36.
[8] P.J. McMurdie, S. Holmes, phyloseq: an R package for reproducible interactive analysis and graphics of microbiome census data, PLOS ONE 8 (4) (2013) e61217, https://doi.org/10.1371/journal.pone.0061217.
[9] G.L. Simpson, P.R. Minchin, M. De Caceres, M.O. Hill, C.J.F. Ter Braak, M.H.H. Stevens, et al., vegan: community ecology package [Internet], 2022 [cited January 14, 2022]. Available from: https://github.com/vegandevs/vegan.
[10] C. Lozupone, R. Knight, UniFrac: a new phylogenetic method for comparing microbial communities, Appl. Environ. Microbiol. 71 (12) (2005) 8228–8235, https://doi.org/10.1128/AEM.71.12.8228-8235.2005.
[11] P. Martinez Arbizu, pairwiseAdonis: pairwise multilevel comparison using adonis [Internet], 2020 [date updated December 30, 2018; date cited February 8, 2023]. Available from: https://github.com/pmartinezarbizu/pairwiseAdonis.
[12] J. Fox, S. Weisberg, An R Companion to Applied Regression [Internet]. Thousand Oaks, CA, USA: Sage [cited February 8, 2023]. Available from: https://socialsciences.mcmaster.ca/jfox/Books/Companion/.
[13] R.V. Lenth, emmeans: estimated Marginal Means, aka Least-Squares Means, R package version 1.5.5-1 [Internet], 2021 [cited February 8, 2023]. Available from: https://cran.r-project.org/package=emmeans.
[14] RStudio Team, RStudio: integrated development for R [Internet], RStudio, PBC, Boston, MA, USA, 2020 [cited February 8, 2023]. Available from: http://www.rstudio.com/.
[15] M.T. Ruel, Operationalizing dietary diversity: a review of measurement issues and research priorities, J. Nutr. 133 (11) (2003) 3911S–3926S.
[16] E.O. Verger, A. Le Port, A. Borderon, G. Bourbon, M. Moursi, M. Savy, et al., Dietary diversity indicators and their associations with dietary adequacy and health outcomes: a systematic scoping review, Adv. Nutr. 12 (5) (2021) 1659–1672.
[17] M.C. Karlsen, G.S. Ellmore, N. McKeown, Seeds—health benefits, barriers to incorporation, and strategies for practitioners in supporting consumption among consumers, Nutr. Today. 51 (1) (2016) 50–59.
[18] R. Ross, I.J. Neeland, S. Yamashita, I. Shai, J. Seidell, P. Magni, et al., Waist circumference as a vital sign in clinical practice: a consensus statement from the IAS and ICCR working group on visceral obesity, Nat. Rev. Endocrinol. 16 (3) (2020) 177–189.
[19] U.S. Department of Agriculture and U.S [Internet], in: Department of Health and Human Services, Dietary Guidelines for Americans, 9th ed, 2020 [cited August 8, 2023]. pp 1–164. Available from: http://www.dietaryguidelines.gov/.
[20] L.A. Ertuglu, A. Demiray, B. Afsar, A. Ortiz, M. Kanbay, The use of Healthy Eating Index 2015 and Healthy Beverage Index for predicting and modifying cardiovascular and renal outcomes, Curr. Nutr. Rep. 11 (3) (2022) 526–535, https://doi.org/10.1007/s13668-022-00415-2.
[21] Y.R. Patel, J.M. Robbins, J.M. Gaziano, L. Djoussé, Mediterranean, DASH, and Alternate Healthy Eating Index dietary patterns and risk of death in the Physicians' Health Study, Nutrients 13 (6) (2021) 1893.