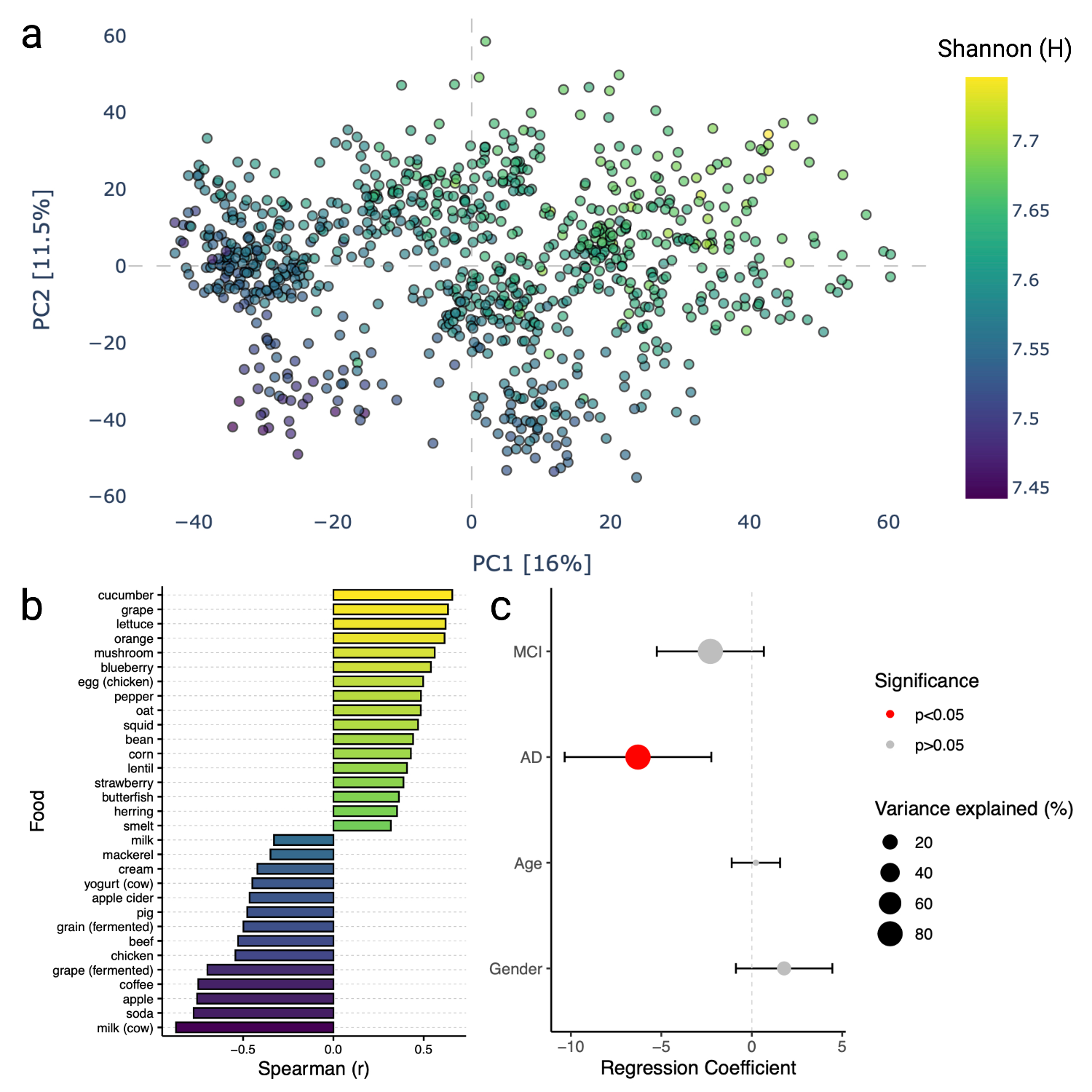
**Supporting information:** **Reference data-set driven metabolomics**

**Authors:** Julia M. Gauglitz, Kiana A. West, Wout Bittremieux, Candace L. Williams, Kelly C. Weldon, Morgan Panitchpakdi, Francesca Di Ottavio, Christine M. Aceves, Elizabeth Brown, Nicole C. Sikora, Alan K. Jarmusch, Cameron Martino, Anupriya Tripathi, Michael J. Meehan, Kathleen Dorrestein, Justin P. Shaffer, Roxana Coras, Fernando Vargas, Lindsay DeRight Goldasich, Tara Schwartz, MacKenzie Bryant, Gregory Humphrey, Abigail J. Johnson, Katharina Spengler, Pedro Belda-Ferre, Edgar Diaz, Daniel McDonald, Qiyun Zhu, Emmanuel O. Elijah, Mingxun Wang, Clarisse Marotz, Kate E. Sprecher, Daniela Vargas-Robles, Dana Withrow, Gail Ackermann, Lourdes Herrera, Barry J. Bradford, Lucas Maciel Mauriz Marques, Juliano Geraldo Amaral, Rodrigo Moreira Silva, Flavio Protasio Veras, Thiago Mattar Cunha, Rene Donizeti Ribeiro Oliveira, Paulo Louzada-Junior, Robert H. Mills, Paulina K. Piotrowski, Stephanie L. Servetas, Sandra M. Da Silva, Christina M. Jones, Nancy J. Lin, Katrice A. Lippa, Scott A. Jackson, Rima Kaddurah Daouk, Douglas Galasko, Parambir S. Dulai, Tatyana I. Kalashnikova, Curt Wittenberg, David J. Gonzalez, Robert Terkeltaub, Megan M. Doty, Jae H. Kim2, Kyung E. Rhee, Julia Beauchamp-Walters, Kenneth P. Wright Jr, Maria Gloria Dominguez-Bello, Mark Manary, Michelli F. Oliveira, Brigid S. Boland, Norberto Peporine Lopes, Monica Guma, Austin D. Swafford, Rachel J. Dutton, Rob Knight, Pieter C. Dorrestein

**Table S1 – The metadata for the foods from Global FoodOmics profiled in this study. This is provided as a separate excel file.**

**Table S2, details of the 28 studies and their public accession number**.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Study** | **Sample Type** | **SSF** | **Num samples** | **Massive ID** |
| **GFOP3500** | Food | N/A | 3527 | MSV000084900 |
| **Sleep and circadian study** | Fecal; Plasma | yes (197) | 98 (F); 371 (P) | MSV000083759 |
| **Centenarian** | Fecal; Plasma | yes (38) | 91 (F); 50 (P) | MSV000084591 |
| **Impact of diet on RA** | Fecal; Plasma | yes (12) | 51 (F); 60 (P) | MSV000084556 |
| **LP Infant** | Fecal; Oral; Skin | yes (58) | 492(F); 461(O); 461(S) | MSV000083462; MSV000083463 |
| **Children with Medical Complexity** | Fecal | yes (24) | 95 | MSV000084610 |
| **American Gut** | Fecal |  | 2123 | MSV000081981 |
| **Fermented food consumption** | Fecal |  | 276 | MSV000081171 |
| **Malawi legume supplement** | Fecal | yes (14) | 1131 | MSV000081486 |
| **Rotarix vaccine response** | Fecal |  | 118 | MSV000084218 |
| **IBD\_1** | Fecal |  | 40 | MSV000082431 |
| **IBD\_individual** | Fecal |  | 5 | MSV000079115 |
| **IBD\_seed** | Fecal |  | 334 | MSV000082221 |
| **IBD\_biobank** | Fecal |  | 95 | MSV000079777 |
| **IBD\_2** | Fecal |  | 206 | MSV000084775 |
| **IBD\_200** | Fecal |  | 203 | MSV000084908 |
| **Alzheimer's disease** | Plasma; CSF |  | 78 (P);  116 (CSF) | MSV000085256 |
| **COVIDovid-19 Brazil** | Plasma | yes (60) | 46 | MSV000085505; MSV000085537 |
| **IBD\_biopsy** | Tissue |  | 135 | MSV000082220 |
| **Gout** | Serum |  | 39 | MSV000084908 |
| **Adult saliva** | Saliva |  | 89 | MSV000083049 |
| **Legume supplementation** | Urine | yes (15) | 5 | MSV000084663 |

****

**Supplementary Figure 1. Diet-associated molecules distinguish Alzheimer’s disease patients from those with normal cognition. a.** Principal component analysis (PCA) of food counts generated from serum samples. Each point is a person in the study. Points are colored by their Shannon diversity index. Each dot is a person in the study. (H). **b.** Correlations using food counts as input between principal component (PC) 1 and relative abundances of foods at level 5 of the food ontology. **c.** Regression coefficients from a linear mixed-effects model represent the associations between PC1 and study group (referenced to normal cognition), age, and gender. Normal cognition: n=286, mild cognitive impairment (MCI): n= 447, Alzheimer’s disease(AD): n=137.

### **Online Methods.**

#### **Data and Code Availability**

The code generated during this study is available on GitHub at <https://github.com/DorresteinLaboratory/GlobalFoodomics>. The following files are available in addition to the Global FoodOmics mzXML files on massive.ucsd.edu under MSV000084900: metadata as a .txt; an image repository with between 1 and 6 images per food item that was sampled; table of FDR-based parameters; full size PDF of sleep restriction and circadian misalignment study, food reference data molecular network (excerpts found in Figure 1). Metadata dictionary can also be accessed here:  
<https://docs.google.com/spreadsheets/d/1Ebn-TgMWEkd_7KOw9TCRvHGPsE7dGjVCr7dg28pwbmM/edit#gid=727944641>

The GNPS based molecular networking analyses used in this study can be accessed online at the following links:

* Sleep and circadian study (MSV000083759; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e0bf255bcb2e492bb0be3be1a691b5fb>; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=6fe434761daf4f9da540cf1fd90b3985>; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9a90bd12f51e453e968656e6458e0da4>)
* Centenarian (MSV000084591; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=8895b6e3445546c4a5bc3a726a920227>; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=981c9a7d39f742bda296d52f856981e5>)
* Impact of diet on RA (MSV000084556; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=0794151fce2c4c18a7a0aa3a09140169>)
* LP Infant (MSV000083462; MSV000083463; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a7b222466ef844e69cdbd9835d2f6c39>; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=c756a9dfb5c34a2a8655f88114edf0a8>; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4a322e640bb644068030949267fb4ea9>)
* Children with Medical Complexity (MSV000084610; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=df24423835a341969342c2086b46275a>)
* American Gut (MSV000081981; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=4884483bcffe4f269819858c3fd4faef>)
* Fermented food consumption (MSV000081171; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=5cca39e0ebab4066a56e41ded48b4466>)
* Malawi legume supplement (MSV000081486; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=93ba727aa9234727a73ae7860b2af3ca>)
* Rotarix vaccine response (MSV000084218; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=08e9b9e048f04ac4b416e574a073e8e6>)
* IBD\_1 (MSV000082431; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=ec08eed8f186430d893c63111409baf4>)
* IBD\_individual (MSV000079115; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=fad746939afd4184975a296436aebfb7>)
* IBD\_seed (MSV000082221; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=907f2e0b7878417dbdb4c83f0df0e83a>)
* IBD\_biobank (MSV000079777; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a79fbd4c96124209adfd0ef84cb56dec>)
* IBD\_2 (MSV000084775; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=07f855658c5342458045032ea70fc526>)
* IBD\_200 (MSV000084908; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=55bef02250d744eb97c6040c379cbfb4>)
* Alzheimer's disease (MSV000085256; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=aac78e9d23b84194ab2f768cb685c636>)
* AD serum (MSV000086270; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=570aacf2244948c7afa590631de5d345>)
* Omnivore vs veg (MSV000086989; https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=74089e95b8df41b2af7c289869dc866f)
* COVID-19 (MSV000085505; MSV000085537; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=9cbcb6b46fe24826bc56c9e893d0bd2b>)
* IBD\_biopsy (MSV000082220; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=a83a279dad154f9ca7b549d40ce117ba>)
* Gout (MSV000084908; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=55bef02250d744eb97c6040c379cbfb4>)
* Adult Saliva (MSV000083049; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=6dd6e5b1cf454d67b8a2b3c151c18f4a>)
* Legume supplementation (MSV000084663; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=93ba727aa9234727a73ae7860b2af3ca>)
* Food only (MSV000084900; <https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d5adba7f67cc402396e9ba7cd85ce52b>)

Networking parameters were set based on the MOLECULAR-LIBRARYSEARCH-FDR workflow on GNPS with the following task IDs:

* GFOP3500: a7bf6cc3f91d466bab923f2268d6f4fc
* Sleep deprivation: b55ab4004ed342d7b4ed1c488e935998
* Sleep study: 78bbfed8574748d1a77dc7c2f1a44d39
* Sleep study\_SSF\_test: b55ab4004ed342d7b4ed1c488e935998
* Centenarian: 265a9553c69e47499cca3de056b43178
* Centenarian\_SSF\_test: 265a9553c69e47499cca3de056b43178
* American Gut: aee5dde3b2f84079a264e68ec981487e
* Fermented food consumption: a44d1b2e1b9d4612974d0b85021675a7
* Malawi legume supplement: de7b55f8adaa4ad9b2a8430e30435bf3
* Children with Medical Complexity: f27243af071b43ab90d846bda959fc1c
* Rotarix vaccine response: a2e02e3f97a54ca08e3866cc60f8d42b
* Impact of diet on RA: 62b8754e761549f3b94ffae83d7ab95a
* LP infant: 532aba2ad3644fadba0e6e7ea063c7ee
* IBD\_1: bb10b1ce90a24f3a9cef1e85e88c3882
* IBD\_biopsy: c4cfda90933b4842a7154f5f2def139d
* IBD\_individual: 3ce8cc636ae944848b4ada322aaf12fe
* IBD\_seed: ebbb715fc605457ba5f7e910b79d6177
* IBD\_biobank: 9465c34cf5444e12b89318b1fb363714
* IBD\_2: 983fa9271136404fb5743b44a6a109f0
* IBD\_200: e5acf5726722486caa897b2b07d402e8
* Impact of diet on RA: 62b8754e761549f3b94ffae83d7ab95a
* Alzheimer's disease: 658103164325425981c097cecba840b0
* AD serum: 67516099b37647f2a9c91f890366bef3
* Omnivore vs vegan: ba974d08cab04f77aaacdb7828baada6
* Gout: a478f419ae824378aa02e5e1b310cad2
* Adult saliva: 32980f95dbd5437aaa9e15d05c7246bb
* LP infant: 8bfbdc1bf38c418fb223306cd42af897
* LP infant: 3e414e13a4394bb78c07f7ca7f4d1be3
* Legume supplementation: 2ca007303b9c4bb3820f392b996eba27
* Alzheimer's disease: 658103164325425981c097cecba840b0
* COVID-19 Brazil: d16eb32276c84bdb9c35c5872e97a986

### **IRB information for the human datasets used in this study and GNPS/MassIVE ID.**

Sleep study (MSV000083759; IRB 15-0282), centenarian (MSV000084591; IRB 180478), Impact of diet on RA (MSV000084556; IRB 161474), LP Infant (MSV000083462; MSV000083463; IRB 151713 UCSD), Children with Medical Complexity (MSV000084610; IRB 161948 UCSD), American Gut (MSV000081981; IRB 141853 UCSD), Fermented food consumption (MSV000081171; IRB 141853 UCSD / published), Malawi legume supplement (MSV000081486; IRB ID #201503171; Washington University Human Studies Committee), Rotarix vaccine response (MSV000084218; IRB is PR-10060 from University of Virginia), IBD\_1 (MSV000082431; IRB # 150675), IBD\_individual (MSV000079115; IRB # 150675), IBD\_seed (MSV000082221; UCSD HRRP 131487), IBD\_biobank (MSV000079777; UCSD HRRP 131487); IBD\_2 (MSV000084775; IRB # 150675), IBD\_200 (MSV000084908; IRB # 150675), Alzheimer's disease (MSV000085256; UCSD IRB # 170957), COVID-19 (MSV000085505; MSV000085537; IRB approval number is 30248420.9.0000.5440 (University of São Paulo, Brazil), IBD\_biopsy (MSV000082220; IRB number is 120025), Gout (MSV000084908; IRB Project #160768X), Adult Saliva (MSV000083049; IRB 150275 UCSD), Legume supplementation (MSV000084663; IRB ID #201905103), NIST Omnivore and Vegan reference data (MSV000086989; de-identified NIST IRB MML-2019-035).

#### **Global FoodOmics reference data**

For the exemplary data set used to highlight RDD metabolomics analysis we created and leveraged the “Global FoodOmics” project (<http://www.globalfoodomics.org>) reference data set. This data set now contains 3,579 food and beverage samples contributed by the community, following in the footsteps of the American Gut and the Earth Microbiome Projects.1,2 The majority of samples were photographed, and a subset were subjected to 16S rRNA profiling (1,511 samples) to characterize the microbial composition, as well as providing information about mitochondria and chloroplast sequences matched by the same primers. Raw and processed 16S rRNA amplicon sequencing data is available at Qiita study #11442 and raw sequence data has been deposited at EBI accession ERP122648. Foods from our Global FoodOmics project were curated according to the Earth Microbiome Project Ontology, the USDA Food Composition Database, a modification to the Food and Nutrient Database for Dietary Studies3,4 <https://ndb.nal.usda.gov/>) and also included a six-level food ontology, as well as information for fermentation or organic status, land or aquatic origin, country of origin, etc.

##### **Sample Collection**

Sampling methodology was developed in order to facilitate sample collection in any environment, from the home, a restaurant, a festival, or in the lab. Initial samples were collected between April 2017 and March 2018. Additional sets of samples were added through Fall 2019. Each sample was assigned a unique number identifier upon sampling, which was used to trace the origin of the sample, and to organize descriptive information about the sample. In addition, when possible, samples were photographed by the participant to create a photographic archive of all samples (uploaded to MassIVE MSV000084900; >4,000 images representing 67% of the samples (2399/3579)). Primarily for the initial data set these images were used as the first point of reference for the collection of ancillary information about the different samples (termed metadata, described in more detail below). The image archive was critical to allow retroactive metadata curation. As the project evolved and the breadth of sample types increased, new categories were added to the metadata, which were then filled in weeks or even months after sample collection.

Samples were frozen at -80oC within 24 h of sample collection, unless otherwise noted in the metadata. Two samples were collected for each food or beverage included in the study. One sample was collected as an archive and directly frozen, and a second sample was collected for extraction. Food samples were collected in a tube prefilled with 1 ml 95% ethanol (Ethyl alcohol (Sigma- Aldrich) and Invitrogen UltraPureTM Distilled Water), as high ethanol concentrations are efficacious at preserving the sample for both DNA and metabolite analyses.5 Samples were collected into 2 ml round bottom microcentrifuge tubes (Qiagen) and weighed prior to freezing. The pre-sample and post-sample weights as well as the weight differences were recorded in the metadata. It was not possible to collect all samples at a given concentration of extraction solvent (ethanol), because sampling was performed in many different environments and is meant to be consistent with future crowd-based community science participation. Therefore the data can be compared qualitatively and not quantitatively, however for certain subsets 50 mg of material were collected.

Additional sets of food samples were added to the core set using the same methods as outlined above when possible. Samples from Venezuela were collected whole in absolute ethanol >=99.8% (Sigma Aldrich) and the extract was processed directly.

The experimental protocol for the sleep restriction and circadian misalignment study has been described previously.6 Meals and food samples were prepared by the Clinical and Translational Research Center Nutrition Core of the Colorado Clinical and Translational Sciences Institute. Food was transported to the research site and refrigerated for the duration of the in-patient study. Individual meals were sampled and stored frozen in ziptop bags. They were stored at -70oC prior to subsampling and LC-MS/MS analysis. Images are contained in a separate Sleep Study folder (MSV000084900).

For several of the human studies we collected data on associated foods (study and region specific foods terms SSF), which were processed according to the same methods as the Global FoodOmics samples. The number of SSF samples per cohort are outlined here: experimental sleep restriction and circadian misalignment (197 samples; 45 are pooled); centenarian (38 individual samples); Malawi legume supplement (14; 2 sample types, several extraction types); children with medical complexity (24 formula samples; 11 exact overlap); RA diet samples (20 individual sample; 2 samples types (stool, plasma), 3 time points)); mother’s milk (58 milk samples); legume supplements (15 individual legume samples; 6 different types).

##### **Community-based science collection**

During the course of sampling, samples were received from over 50 different individuals in California as well as from different states as well as countries (such as Malawi, Venezuela, Italy and Brazil). Contributions from individuals ranged from produce from home gardens, home fermented products (yogurt, kombucha, sauerkraut), meat and dairy from private farms, to items individuals had purchased that were of interest to them.

We were also directly invited to sample at local stores and organizations, including Venissimo cheese, Good Neighbor Gardens, and the San Diego Zoo and San Diego Zoo Safari Park, as well as local supermarkets such as Sprouts Farmers Market, Whole Foods Market, and Ralphs. We were invited by San Diego Fermenter’s Club founder Austin Durant to the San Diego Fermenter’s Club meeting and sampled from multiple vendors at both the Oregon Fermentation Festival in 2017 as well as the San Diego Fermentation Festival in 2018. We also received citrus samples from a farm at the US-Mexico border, with visibly dark skin due to air pollution, a particular concern of the farmer. Other sampling occurred in conjunction with study design, as was the case for the Rheumatoid arthritis cohort and the COVID-19 study. In total we engaged with a broad range of individuals, organizations, businesses and scientists, to generate this dataset of 3579 samples (and continues to be expanded). A predominance of foods included in this initial dataset were sampled and/or purchased in California, leaving room for much further expansion and the inclusion of a crowd-sourced community science initiative to expand the array of samples.

The sample set contains a broad set of simple foods including fruits, vegetables, grains/legumes, as well as raw meat and fish, which build the foundation of many food products. In addition, we have 1133 fermented samples. This subcategorization of foods is made possible by the metadata collected on these samples, described in the Metadata Curation section. The breadth of samples included in the dataset necessitated careful collation and a range of information about the samples, resulting in 157 different metadata categories to describe various aspects of these food and beverage samples (Table S1).

The foods, although primary consumed in the US, could be traced to originate from over 50 different countries or territories of origin reflecting the global distribution of food (Argentina, Australia, Austria, Belgium, Bolivia, Brazil, Canada, Chile, China, Colombia, Croatia, Ecuador, England, Ethiopia, France, Germany, Greece, Guatemala, Haiti, Holland/Netherlands, India, Indonesia, Ireland, Israel, Italy/Sardinia, Japan, Kenya, Korea, Madagascar, Malawi, Mexico, New Zealand, Nilgiri, Peru, Philippines, Poland, Serbia, Portugal, Russia, Scotland, South Africa, Spain, Switzerland, Taiwan, Thailand, Trinidad & Tobago, Turkey, UK, USA/Puerto Rico, Vietnam, Venezuela; some are labeled by continent such as US, EU or South America.

##### **Metadata Curation**

Detailed information about each sample was captured in the form of metadata. There are 157 metadata fields available for each food. The metadata are in the form of an array, where each row represents one sample and each column captures unique information about the sample (See Supplementary Information for Metadata File, as well as metadata on Massive MSV000084900). This matrix allows for the categorization of foods by various different attributes and links these attributes to the sample numbers, the data files (.mzXML filename), as well as the 16S sequence information on Qiita (sample\_name). The initial metadata categories captured included sample description, sample number, location sample was collected, the weight of the sample (pre-sample, post-sample, sample weight), the day it was collected, and whether an image had been taken and renamed to match the sample number and archived in the image repository. The initial 9 categories captured minimal information and allowed tracking of information about the sample.

During the process of sample collection, the diversity of the samples being collected necessitated the addition of columns to capture more information about the samples and to be able to categorize them and compare different attributes. These columns grew to capture highly detailed information about each sample, for example whether the sample was organic, if it was raw or cooked, if it was washed before sampling, or for cheese samples whether it is the rind or the curd, etc. As columns were added, the initial columns and the image repository were used to trace back information.

###### **Classification scheme**

Various classifiers are used to describe foods, however we were unable to find an established scheme able to capture the diversity of samples, as well as distill the metadata down into a manageable number of categories to distinguish differences between the metabolomes of different food classes. We therefore categorized the foods by sample\_type, which captured whether the sample was a food, beverage, or other item (for example supplements) and then expanded and shaped a unique categorization which takes into account the species and botanical definitions of foods. The sample\_type categories range from sample\_type\_land\_aquatic, to differentiate items sourced from different physical environments, sample\_type\_common, which allows for representation of a particular food group which was not otherwise captured in the metadata, such as zoo food or candy. The sample\_type groups also include a hierarchy from group1 to group6 (Levels 1 through 5 are referenced in this manuscript), specific to foods and groupB1 through groupB3 which contain beverage specific information (alcoholic [binary], carbonated [binary], type of beverage [such as red wine, kefir, soda, etc.]).

###### **Complex samples**

The above classification scheme gave sufficiently detailed information about simple foods (ones that have only one ingredient and could thus be filled out to the last group level, such as red cherry tomato). Complex foods contain not only multiple ingredients, but include highly processed foods purchased with ingredient lists as well as home cooked or restaurant meals. These foods have a higher variability of information known about them. When available, the top 6 ingredients are captured in individual metadata categories, with a seventh ingredient field which contains the remainder of the ingredients. However, the order of ingredients does not always clearly reflect the type of food and some constituents that may be of interest, such as tree nuts which may only be found in trace quantities. The sample\_type\_common category captured some of the information about the type of sample (candy), however to have a tangible classification of different ingredient types, we generated a specific complex food ontology based on the known presence of common categories (corn, dairy\*, egg\*, fruit, fungi, fish\*, shellfish\*, meat, peanut\*, seaweed, soy\*, tree nut\*, vegetable/herb, wheat\* (\*designates known food allergen)). These categories reflect the main food groups and some of the most common allergens (US FDA Food Allergen Labeling And Consumer Protection Act of 2004)7, items which are of interest when correlating food metabolome data with other datasets, such as human fecal material (where the foods eaten are known or unknown).

###### **Fermented foods**

Preservation and processing methods are included in the metadata. However, due to the potential importance of fermentation in the alteration of the food metabolome, and the potential health benefits that have been ascribed to fermented foods, several categories were included to highlight this feature: fermented or not, whether it contains live active cultures, whether it contains chocolate (which then was cross checked with the fermented category, as chocolate is a fermented food). The list of fermented foods crosses many of our sample types as it includes fermented dairy (yogurt, cheese), fermented meat/fish (salami, fish sauce), fermented vegetables (kimchi, sauerkraut), fermented fruit (chocolate, coffee, apple), and fermented grains/legumes (bread, tempeh).

###### **Food specific categories**

Certain individual food categories also necessitated creation of specific categorization. For example, cheeses have the specific categories cheese\_part (curd vs. rind), cheese\_type (washed, blue, etc), and cheese\_texture (soft, semi-soft, semi-hard, hard). Particularly for raw plant products, such as fruits, vegetables, grains which form the basis for many food ingredients, we captured botanical information: botanical\_anatomy (fruit, leaf, tuber, seed, etc.), botanical\_genus, and botanical\_genus\_species (when known). Tea samples have tea quality and tea type as distinct categories.

###### **Metadata for Cross-study Comparison**

To facilitate cross study comparison, we included the Earth Microbiome Project ontology: empo\_1 (level 1: Free-living, Host-associated, Control, or Unknown), empo\_2 (level 2: Saline, Non-saline, Animal, Plant, or Fungus), and empo\_3 (level 3: most specific habitat name) [[http://earthmicrobiome.org/protocols-and-standards/empo/](http://www.earthmicrobiome.org/protocols-and-standards/empo/)]. Wherever possible we linked foods to food identifiers or created identifiers and categories that built upon the existing framework as defined by the U.S. Department of Agriculture's Food and Nutrient Database for Dietary Studies 2011-2012 (FNDDS) food grouping scheme.8

**Metabolite Extraction**

Homogenized samples (in ethanol) were incubated for 40 min at -20°C and centrifuged (Eppendorf centrifuge 5418, Hamburg, Germany) at 20,000 rpm for 15 min at 4°C. 400 μL of supernatant were transferred to a 96-well deep well plate and dried by centrifugal evaporation (Labconco Acid-Resistant Centrivap Concentrator, Missouri, USA). Dried extracts were reconstituted in 150 μL of resuspension solution (50% methanol with 2 μM sulfadimethoxine), then vortexed for 2 min and sonicated for 5 min in a water bath (Branson 5510, Connecticut, USA). Resuspended extracts were then centrifuged for 15 min at 20,000 rpm and 4°C (Thermo SORVALL LEGEND RT, Germany) and transferred to a 96-well shallow well plate, and diluted either 5x or 10x to avoid saturating the MS detector.

##### **Liquid Chromatography - Mass Spectrometry**

Food extracts were analyzed using an UltiMate 3000 UHPLC system (Thermo Scientific, Waltham, Ma) equipped with a reverse phase C18 column, prepended with a guard cartridge (Kinetex, 100 x 2.1 mm, 1.7 μm particles size, 100 Å pore size; Phenomenex, Torrance, CA, USA), at a column compartment temperature of 40°C. Samples were chromatographically separated with a constant flow rate of 0.5 mL / min using the following gradient: 1.5 min isocratic at 5% B, up to 100% B in 8 min, 3 min isocratic at 100% B, back to 5% B in 0.5 min and then 1.5 min isocratic at 5% B (A: H2O + 0.1% formic acid; B: Acetonitrile (ACN) + 0.1% formic acid (LC-MS grade solvents, Fisher Chemical, Hampton, United States)).

The UHPLC system was coupled to a Maxis Q-TOF Impact II mass spectrometer (Bruker Daltonics, Bremen, Germany) equipped with an electrospray ionization source. MS spectra were acquired in positive ionization mode using Data Dependent Acquisition (DDA) with a mass range of *m/z* 50–1500. The instrument was externally calibrated two times per day to 1.0 ppm mass accuracy using ESI-L Low Concentration Tuning Mix (Agilent Technologies, Waldbronn, Germany). Hexakis (m/z 622.029509; (1H, 1H, 2H difluoroethoxy)phosphazene (Synquest Laboratories, Alachua, FL)) was used for lock mass correction. MS/MS spectra were acquired for the top 5 ions in each MS1 spectrum, with active exclusion after 2 spectra (maintained for 30 seconds). Known contaminants as well as lockmass values commonly used with this instrument were added to an exclusion list (*m/z* values listed): 144.49–145.49; 621.00–624.10; 643.80–646.00; 659.78–662.00; 921.0–925.00; 943.80–946.00; 959.80–962.00.

Raw high resolution mass spectrometry data files were converted to open source .mzXML format using Bruker DataAnalysis software after lock mass correction (*m/z* 622.0290). Raw data files as well as converted .mzXML files were uploaded to MassIVE (publicly available under unique identifier MSV000084900) and further analyzed on Global Natural Product Social Molecular Networking (GNPS) (https://gnps.ucsd.edu), as described below.

##### **MS2 Data Processing**

##### **FDR estimation**

False discovery rate (FDR) estimation was calculated using Passatutto analysis workflow in GNPS.9,10 FDR estimation was used to determine the cosine value required with a minimum of 5 matched peaks to achieve an FDR of 1%. See the Data availability section for accession information.

###### **Molecular networking using GNPS**

Molecular networking analysis and library search were performed using GNPS classical molecular networking release\_18.10 3579 .mzXML data files (available at MassIVE ID MSV000084900) were included in the analysis. The data were filtered by removing all MS/MS peaks within +/- 17 *m/z* of the precursor *m/z*. MS/MS spectra were window filtered by choosing only the top 5 peaks in the +/- 50 *m/z* window throughout the spectrum. The data were then clustered with MS-Cluster with a parent mass tolerance of 0.02 *m/z* and an MS/MS fragment ion tolerance of 0.02 *m/z* to create consensus spectra. Further, consensus spectra that contained less than 2 spectra were discarded. A network was then created where edges were filtered to have a cosine score above 0.65 (slight variation per study based on FDR calculation) and more than 5 matched peaks. Further, edges between two nodes were kept in the network if and only if each of the nodes appeared in each other's respective top 10 most similar nodes. The spectra in the network were then searched against the GNPS spectral libraries. The library spectra were filtered in the same manner as the input data. All matches kept between network spectra and library spectra were required to have the same cosine score and minimum matched peaks as for library search. Version release 18 was used to process all studies with the exception of the COVID-19 dataset, which was processed with identical methods and version 23.

Molecular networking analysis utilizes a spectral library of 150,633 public reference spectra that are used by the GNPS analysis infrastructure for annotation of public data which presently includes 29 spectral libraries, including from the three MassBanks (Japan, EU and North America)11, HMDB12, ReSpect13, NIH natural product libraries14, PNNL lipid library 16 24, Bruker/Sumner, FDA libraries, Gates Malaria library, EMBL library, as well as many other GNPS contributed libraries (<https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp>,10) and the commercial NIST17 library (CID portion only). Molecular networks were visualized in the GNPS browser as well as with the freely available program Cytoscape (version 3.5.1).17

###### **Interpreted spectral rate calculation**

The levels of interpretation are delineated as follows: A spectral match between an MS/MS spectrum from human or food data with a library spectrum constitutes a *molecular ID* and determines the initial percent of interpreted spectra, which is also equivalent to the annotation rate of the dataset. A spectral match between MS/MS spectra in human and reference samples (by performing molecular networking of the datasets together and identifying nodes with overlap between the two groups) indicates a *potential source*. Matches between human and food data therefore implicate food as the potential source of the molecule. Food reference data are referred to in two main categories: the Global FoodOmics dataset (GFOP; broad range of foods and beverages) and study specific food (SSF; foods and/or beverages known to be consumed by some participants). The last level of interpretation is based on connectivity within a molecular family, which allows us to infer *structural relatedness* or *possible metabolism* of food derived compounds.

Food reference data and human data were organized into separate groups in the molecular networking analysis. The annotation and interpreted spectral rates were calculated using R (3.6.3) and the *tidyr* and *dplyr* packages. We first calculated percent annotation rate, or molecular ID, for all studies (stool, plasma, etc.) (e.g. # of stool nodes with a molecular ID / total # of stool nodes). Spectral matches between food reference data and human MS data (overlap between the two groups) provides the next level of information, referred to as the interpreted spectral rate (e.g. # of nodes found in food and stool data / total # of stool nodes), indicating a potential food source.

For molecules without annotations to reference libraries, we wanted to measure the potential to explain their presence using molecular networking. By removing single loops in each dataset and comparing metabolites that shared a component index with an annotated compound, we were able to identify molecules that belong to the same molecular family to infer their potential classification, and calculate the interpreted spectral rate by dividing unannotated molecules that network with annotated ones by total metabolites within each sample type. Overlap between sample types was again assessed to understand contributions due to co-networking of molecules across sample types, increasing our ability to explain unannotated molecules found in our datasets. Visualizations were generated using *graphics* and *beeswarm* packages, and significant differences were calculated using Welch’s *t-*tests (*stats::*t.test*),* Welch’s *F*-test (*onewaytests::*welch.test), and Games-Howell (*rstatix::*games\_howell\_test) for multiple comparisons, as appropriate, with multiple comparisons correction using Tukey’s method. All data are expressed as the mean ± standard error and considered significant if *P <* 0.05 unless otherwise stated.

For example, for GNPS molecular networking analyses test datasets were consistently placed in group 1 (G1) (and G2 for paired datasets, such as stool and plasma) and Global FoodOmics data were placed in group 4 (G4). SSFs were consistently placed in G3 when used. The common nodes between G1 and G4 represent the overlap and potential enhancement of information, directly from the reference dataset. The improvement is thus measured by the difference in the overlap of G1 and G4 divided by the total nodes in G1 versus the number of annotations in G1 divided by the total nodes in G1. The “propagation” refers to the counting of nodes within connected components in molecular families which capture three types of additional information: 1) unnannotated compounds found only in G1 that network with an annotated compound found in G4 (could be an annotated molecule observed only in G4 or in G4 and G1), 2) unnannoted compounds found only in G1, but in the same molecular family with an unannotated food compound (G4), or 3) unnannotated compounds found only in G1, but in the same molecular family with an annotated food compound (G4). The increase shown for Total is taking into account the number of unique nodes from the three different types of molecular connectivity. The second is the largest contributor.

#### **Metadata inference - food count generation**

Food counts were calculated as the number of consensus nodes in the molecular networking results that match to food samples. Consensus nodes were required to match to all of the relevant experiment groups (sample type, GFOP, optionally SSFs) and not match to any of the other experiment groups. All source file names corresponding to the filtered consensus nodes were matched to the GFOP file names and metadata to derive counts of the foods at different levels of the food hierarchy. Infrequent food types that occurred less often than water (presumed blank) were removed to filter out sporadic random matches. For the flow diagram the food counts for the complete datasets were calculated at different levels of the metadata hierarchy. Flow diagrams were generated in Python (version 3.8) using Pandas (version 0.25.3), NumPy (version 1.18.1), and floweaver (version 2.0.0a5).17-19

RDD metabolomics based food counts does come with caveats to consider. First, as it employs a database, the depth, breadth, and type of database must be taken into account when interpreting the output. Expanding the general food database with regional foods increased the number of matched spectra, whereas the participant diet diaries still contained foods not yet captured in the food database. Community contributions to expand the database, with high-quality associated metadata to achieve a more complete coverage, will ultimately eliminate this issue. Another consideration is that a molecule could be produced by humans but also be part of different diet sources (i.e. cholesterol produced by the human body versus consumed from meat). However, the RDD method does not rely on a single MS/MS match, but aggregates tens to thousands of matches into signatures that point to specific food categories. The overlap of such matches still contributes to the formulation of a hypothesis that the observed MS/MS features from human data might originate from the reference data as source.

#### **Diet information from the NIST omnivore and vegan reference data.**

Human whole stool was obtained from volunteer donors by The BioCollective (Denver, CO, USA). The samples consisted of whole stool from vegan and omnivore donors (4 donors per cohort) homogenized in deionized water and aliquoted into 1 mL vials. The samples were stored in aqueous and lyophilized conditions at -80°C.

A feature table detailing the number of MS/MS matches between each fecal sample and each food contained in the reference database was generated. Food counts were modelled by principal component analysis (PCA) using the “mixOmics”[(Rohart et al. 2017)](https://paperpile.com/c/KWCSWt/CGeX) package in R. Counts were aggregated for specific food categories (dairy, meat, seafood, legume, fleshy fruit, vegetable/herb) known to be preferentially consumed in either diet. Differences in sum-normalized counts for each food category between omnivore and vegan samples were assessed by Wilcox test.

#### **Diet variation in Alzheimer’s disease patients**

As described above, a feature table was generated based on MS/MS matches between each serum sample and each reference food, then variation in diet readouts was assessed by PCA. Diet alpha-diversity was calculated using the Shannon Index (R package *vegan*). Additionally, feature tables at different levels (L3, L4, and L5) of the food ontology were generated and counts were sum normalized. Correlations (Spearman) between each food category and PC1 were calculated (R package *Hmisc*) to determine dietary patterns. Associations between dietary patterns (PC1) and study group, age, and gender were evaluated using a linear mixed-effects model (R package *lme4*) to control for the random effect of running samples on different plates. The Kenward-Roger approximate *F*-test, as implemented in *pbkrtest*, was used to assess the significance of each fixed effect in the model.

#### **Dataset descriptions**

All human datasets were processed by LC-MS/MS on high resolution mass spectrometers, in positive ionization mode and contained between 5 and 2123 samples, representing multiple different biofluids and tissues (Table S1).

Data were collected for the following studies using a QTOF mass spectrometer and similar methods as those outlined above: American Gut (MSV000081981), Children with Medical Complexity (MSV000084610), Rotarix vaccine response (MSV000084218), Malawi legume supplement (MSV000081486), IBD\_1 (MSV000082431), IBD\_individual (MSV000079115), Fermented food consumption (MSV000081171)20, the Sleep restriction and circadian misalignment (MSV000083759; IRB 15-0282), centenarian (MSV000084591; IRB 180478), and Legume supplementation (MSV000084663). The LP Infant (MSV000083462; MSV000083463), IBD\_seed (MSV000082221), IBD\_biobank (MSV000079777), IBD\_2 (MSV000084775), IBD\_200 (MSV000084908) 30, IBD\_biopsy (MSV000082220), Gout (MSV000084908), Adult Saliva (MSV000083049).

The datasets for the impact of diet on RA (MSV000084556) and Alzheimer's disease (MSV000085256) were collected with similar methods on a Q-exactive Orbitrap mass spectrometer (Thermo Scientific). The Alzheimer samples include Alzheimer’s Disease and elderly controls, and were drawn in the early morning after fasting for at least 6 h.

The food and plasma data for the COVID-19 study (MSV000085505; MSV000085537) were collected at the University of São Paulo, Brazil, as described below: Plasma samples were collected from patients with laboratory confirmed COVID-19 who were admitted to the Special Unit for the Treatment of Infectious Diseases (UETDI) at the General Hospital of the Medical School of Ribeirão Preto (HC-FMRP-USP). Previously, clarifications to patients occurred both orally and in writing, based on the printed text of the Free and Informed Consent Form, which contained the general proposal of the study, the procedures for obtaining the samples, the risks and benefits. In addition, they were assured about confidentiality of their name, personal data and the possibility of giving up their participation at any time. Following the signature, patients received a copy of the informed consent form. The following were included: 1) Patients diagnosed with COVID-19 in moderate, severe or critical forms and in need of hospital treatment; 2) Over 18 years old; 3) At least 50 kg of body weight; 4) Admission electrocardiogram without changes in rhythm and with QT interval <450 ms; 5) normal serum levels of Ca2+ and K+; 6) If a woman, between 18 and 50 years old, negative β-HCG test on admission. Patients were excluded who: 1) have the mild forms of SARS-CoV-2; 2) pregnant; 3) unable to understand the information contained in the Free and Informed Consent Form (ICF).

Sample preparation: For the COVID-19 plasma samples, aliquots of 20 μL were transferred to eppendorf tubes and 120 μL of cold extracting solution, MeOH: MeCN (1: 1, v/v) was added. After orbital shaking for 1 min (Gehaka AV-2 Shaker, São Paulo, Brazil), the samples were left at -20oC for 30 minutes and then centrifuged for 10 min at 20000 × g at 4oC (Centrifuge Boeco Germany M-240R, Germany). An aliquot of the organic phase (120 μL) was transferred to another eppendorf tube and evaporated to dryness in a rotary vacuum concentrator for 60 min, at 30oC (Analitica, Christ RVC2-18, São Paulo). The residues were resuspended in 80 μL of H2O and centrifuged (10 min, 5000 ×g, 4oC), an aliquot of 5 μL was injected.

Mass spectrometry data collection plasma sample extracts were chromatographically separated with an HPLC (Shimadzu, Tokyo, Japan), coupled with a micrOTOF-Q II mass spectrometer (Bruker Daltonics, Boston, MA, USA) equipped with an ESI source and a quadrupole-time of flight analyzer (qTOF, Bruker Daltonics Inc., Billerica, MA, USA). For chromatographic analyses, we employed a Kinetex C18 column (1.7 µm, 100 × 2.1 mm) (Phenomenex, Torrance, CA, USA) kept at 40oC, with a flow rate of 0.3 mL/min. A linear gradient was applied: 0-1.5 min isocratic at 5% B, 1.5-9.5 min 100% B, 9.5-12 min isocratic at 100% B, 12-12.5 min 5% B, 12.5-14 min 5% B; where mobile phase A is water with 0.1% formic acid (v/v) and phase B is acetonitrile 0.1% formic acid (v/v) (LC-MS grade solvents). The MS data were acquired in positive mode using an MS range of *m/z* 50–1500. The equipment was calibrated with trifluoroacetic acid (TFA) every day, and internally during each run. The MS parameters were established as follows: end plate offset, 450 V; capillary voltage, 3500 V; nebulizer gas pressure, 4.0 Bar; dry gas flow, 9 L/min; dry temperature, 220oC.

For data dependent acquisition the five most abundant ions per MS1 scan were fragmented and the spectra collected. MS/MS active exclusion was set after 2 spectra and released after 30 seconds. A fragmentation exclusion list was set: *m/z* 144.49-145.49; 621.00-624.10; 643.80- 646.00; 659.78-662.00; 921.0-925.00; 943.80-946.00; 959.80-962.00 to exclude known contaminants and infused lockmass compounds. A process blank was run every 5 samples; 5 µL of a standard mix [Paclitaxel 1 mg L-1, and Diazepam 1 mg L-1] (Sigma-Aldrich, Saint Louis, Missouri, US) in 50% MeOH (LC-MS grade solvents) was injected every 5 samples. All MS data were analyzed with Bruker Compass DataAnalysis 4.3 software (Bruker Daltonics, Boston, MA, USA).

A metadata file was created grouping all available clinical information from patients with laboratory confirmed COVID-19 and essential analysis specifications. The MS/MS data were calibrated with an internal standard (TFA), converted to mzXML files using MSConvert from the ProteoWizard software [(Chambers et al. 2012)](https://paperpile.com/c/KWCSWt/j03Y4) and then uploaded into the Global Natural Products Social Molecular Networking web-platform (<https://gnps.ucsd.edu/>). All MS data (.mzXML files) and metadata (.txt file) are publically available via GNPS/MassIVE (<https://massive.ucsd.edu/>) under accession number MSV000085373.

**SI References:**

1. Thompson, Luke R., Jon G. Sanders, Daniel McDonald, Amnon Amir, Joshua Ladau, Kenneth J. Locey, Robert J. Prill, et al. 2017. "A Communal Catalogue Reveals Earth's Multiscale Microbial Diversity." Nature 551 (7681): 457-63.
2. McDonald, Daniel, Embriette Hyde, Justine W. Debelius, James T. Morton, Antonio Gonzalez, Gail Ackermann, Alexander A. Aksenov, et al. 2018. “American Gut: An Open Platform for Citizen Science Microbiome Research.” mSystems 3 (3).
3. Sicherer, Scott H., and Hugh A. Sampson. 2006. "9. Food Allergy." The Journal of Allergy and Clinical Immunology 117 (2 Suppl Mini-Primer): S470-75.
4. Martin, Carrie L., Janice B. Montville, Lois C. Steinfeldt, Grace Omolewa-Tomobi, Kaushalya Y. Heendeniya, Meghan E. Adler, and Alanna J. Moshfegh. 2012. "USDA Food and Nutrient Database for Dietary Studies 2011--2012: Documentation and User Guide. Beltsville, MD: US Department of Agriculture." Agricultural Research Service, USDA Food Surveys Research Group.
5. Song, Se Jin, Amnon Amir, Jessica L. Metcalf, Katherine R. Amato, Zhenjiang Zech Xu, Greg Humphrey, and Rob Knight. 2016. "Preservation Methods Differ in Fecal Microbiome Stability, Affecting Suitability for Field Studies." mSystems 1 (3).
6. Sprecher, Kate E., Hannah K. Ritchie, Tina M. Burke, Christopher M. Depner, Alexandra N. Smits, Pieter C. Dorrestein, Monika Fleshner, et al. 2019. “Trait-like Vulnerability of Higher-Order Cognition and Ability to Maintain Wakefulness during Combined Sleep Restriction and Circadian Misalignment.” Sleep 42 (8).
7. https://www.fda.gov/food/food-allergensgluten-free-guidance-documents-regulatory-information/food-allergen-labeling-and-consumer-protection-act-2004-falcpa
8. https://www.ars.usda.gov/ARSUserFiles/80400530/pdf/fndds/fndds\_2011\_2012\_doc.pdf
9. Scheubert K, Hufsky F, Petras D, Wang M, Nothias LF, Dührkop K, Bandeira N, Dorrestein PC, Böcker S. Significance estimation for large scale metabolomics annotations by spectral matching. Nat Commun. 2017 Nov 14;8(1):1494.
10. Wang, Mingxun, Jeremy J. Carver, Vanessa V. Phelan, Laura M. Sanchez, Neha Garg, Yao Peng, Don Duy Nguyen, et al. 2016. “Sharing and Community Curation of Mass Spectrometry Data with Global Natural Products Social Molecular Networking.” Nature Biotechnology 34 (8): 828–37.
11. Horai, Hisayuki, Masanori Arita, Shigehiko Kanaya, Yoshito Nihei, Tasuku Ikeda, Kazuhiro Suwa, Yuya Ojima, et al. 2010. "MassBank: A Public Repository for Sharing Mass Spectral Data for Life Sciences." Journal of Mass Spectrometry: JMS 45 (7): 703-14.
12. Wishart, David S., Yannick Djoumbou Feunang, Ana Marcu, An Chi Guo, Kevin Liang, Rosavazquez-Fresno, Tanvir Sajed, et al. 2018. "HMDB 4.0: The Human Metabolome Database for 2018." Nucleic Acids Research 46 (D1): D608-17.
13. Sawada, Yuji, Ryo Nakabayashi, Yutaka Yamada, Makoto Suzuki, Muneo Sato, Akane Sakata, Kenji Akiyama, et al. 2012. "RIKEN Tandem Mass Spectral Database (ReSpect) for Phytochemicals: A Plant-Specific MS/MS-Based Data Resource and Database." Phytochemistry 82 (October): 38-45.
14. Huang, Ruili, Hu Zhu, Paul Shinn, Deborah Ngan, Lin Ye, Ashish Thakur, Gurmit Grewal, et al. 2019. “The NCATS Pharmaceutical Collection: A 10-Year Update.” Drug Discovery Today 24 (12): 2341–49.
15. Kyle, Jennifer E., Kevin L. Crowell, Cameron P. Casey, Grant M. Fujimoto, Sangtae Kim, Sydney E. Dautel, Richard D. Smith, Samuel H. Payne, and Thomas O. Metz. 2017. "LIQUID: An-Open Source Software for Identifying Lipids in LC-MS/MS-Based Lipidomics Data." Bioinformatics 33 (11): 1744-46.
16. Shannon, Paul, Andrew Markiel, Owen Ozier, Nitin S. Baliga, Jonathan T. Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, and Trey Ideker. 2003. "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks." Genome Research 13 (11): 2498-2504.
17. McKinney, Wes. 2010. “Data Structures for Statistical Computing in Python.” In Proceedings of the 9th Python in Science Conference, 56–61. Proceedings of the Python in Science Conference. SciPy.
18. Walt, S. van der, S. C. Colbert, and G. Varoquaux. 2011. “The NumPy Array: A Structure for Efficient Numerical Computation.” Computing in Science Engineering 13 (2): 22–30.
19. Lupton, R. C., and J. M. Allwood. 2017. “Hybrid Sankey Diagrams: Visual Analysis of Multidimensional Data for Understanding Resource Use.” Resources, Conservation and Recycling 124 (September): 141–51.
20. Taylor, Bryn C., Franck Lejzerowicz, Marion Poirel, Justin P. Shaffer, Lingjing Jiang, Alexander Aksenov, Nicole Litwin, et al. 2020. "Consumption of Fermented Foods Is Associated with Systematic Differences in the Gut Microbiome and Metabolome." mSystems 5 (2).