nature biotechnology

BRIEF COMMUNICATION

https://doi.org/10.1038/s41587-022-01368-1

Enhancing untargeted metabolomics using metadata-based source annotation

Julia M. Gauglitz[1,2,36], Kiana A. West[1,2,36], Wout Bittremieux[1,2,36], Candace L. Williams[3], Kelly C. Weldon[1,2,4], Morgan Panitchpakdi[1,2], Francesca Di Ottavio[1], Christine M. Aceves[1,2], Elizabeth Brown[2,5], Nicole C. Sikora[1,2], Alan K. Jarmusch[1,2], Cameron Martino[4,6,7], Anupriya Tripathi[2,5,6], Michael J. Meehan[1,2], Kathleen Dorrestein[1,2], Justin P. Shaffer[6], Roxana Coras[8], Fernando Vargas[1,2,5], Lindsay DeRight Goldasich[6], Tara Schwartz[6], MacKenzie Bryant[6], Gregory Humphrey[6], Abigail J. Johnson[9], Katharina Spengler[1], Pedro Belda-Ferre[4,6], Edgar Diaz[6], Daniel McDonald[6], Qiyun Zhu[6], Emmanuel O. Elijah[1,2], Mingxun Wang[1,2], Clarisse Marotz[6], Kate E. Sprecher[10,11], Daniela Vargas-Robles[12], Dana Withrow[10], Gail Ackermann[6], Lourdes Herrera[13], Barry J. Bradford[14], Lucas Maciel Mauriz Marques[15], Juliano Geraldo Amaral[16], Rodrigo Moreira Silva[17], Flavio Protasio Veras[15], Thiago Mattar Cunha[15], Rene Donizeti Ribeiro Oliveira[18], Paulo Louzada-Junior[18], Robert H. Mills[1,2,6,19], Paulina K. Piotrowski[20], Stephanie L. Servetas[20], Sandra M. Da Silva[20], Christina M. Jones[20], Nancy J. Lin[20], Katrice A. Lippa[20], Scott A. Jackson[20], Rima Kaddurah Daouk[21,22,23], Douglas Galasko[24], Parambir S. Dulai[25], Tatyana I. Kalashnikova[26], Curt Wittenberg[26], Robert Terkeltaub[8,27], Megan M. Doty[6,28], Jae H. Kim[29], Kyung E. Rhee[6], Julia Beauchamp-Walters[30], Kenneth P. Wright Jr[10], Maria Gloria Dominguez-Bello[31], Mark Manary[32], Michelli F. Oliveira[33], Brigid S. Boland[25], Norberto Peporine Lopes[17], Monica Guma[8], Austin D. Swafford[4], Rachel J. Dutton[5], Rob Knight[4,6,33,34,35] ✉ and Pieter C. Dorrestein[1,2,4,6,19] ✉

Human untargeted metabolomics studies annotate only ~10% of molecular features. We introduce reference-data-driven analysis to match metabolomics tandem mass spectrometry (MS/MS) data against metadata-annotated source data as a pseudo-MS/MS reference library. Applying this approach to food source data, we show that it increases MS/MS spectral usage 5.1-fold over conventional structural MS/MS library matches and allows empirical assessment of dietary patterns from untargeted data.

Complex sequence data from metagenomic (see Box 1 for definition of terms) or metatranscriptomic experiments require for interpretation both databases of curated genes and reference data, such as whole genomes or other sequence data with carefully curated metadata (developmental stage, tissue location, phenotype, etc.)[1–4]. Such reference data-driven (RDD) analysis increases understanding of complex communities by using matches between genes or transcripts of known and unknown origin. The RDD strategy is essential for the successful analysis of most metatranscriptomics or metagenomics data. By analogy, interpreting liquid chromatography–tandem mass spectromtery (LC–MS/MS)-based untargeted metabolomics data is performed by searching structural MS/MS libraries. However, leveraging reference data with curated and structured controlled vocabulary metadata to improve insights obtainable from untargeted MS/MS-based metabolomics is not yet done.

RDD analysis uses not only annotated MS/MS-spectra but also all unannotated spectra. The gas chromatography–mass spectrometry (GC–MS) BinBase resource has made a step in the direction of RDD. With BinBase one can annotate if a spectrum match has been observed in a non-public GC–MS dataset. However, the metadata is not well controlled and lacks the ability to add contextualized metadata[5,6]. In addition, as we have previously demonstrated, using structural annotations, the source can be determined by literature mining[7]. However, owing to the above mentioned limitations and/or inability to link related spectra in the case of metabolism, the above strategies to annotate unknowns cannot be used to systematically to interpret the source information at the dataset level. We therefore introduce the RDD approach for metabolomics (Fig. 1), followed by a use case demonstrating empirical food readouts from untargeted human data (Fig. 2).

Untargeted MS/MS-based metabolomics experiments involve searching MS/MS structural libraries since the late 1970's[8,9], or, more recently, for investigating the distribution of a MS/MS spectrum across public untargeted data[10]. Instead of only leveraging a single MS/MS spectrum to obtain an annotation, RDD metabolomics uses all MS/MS spectra from untargeted metabolomics files, which con-

A full list of affiliations appears at the end of the paper.

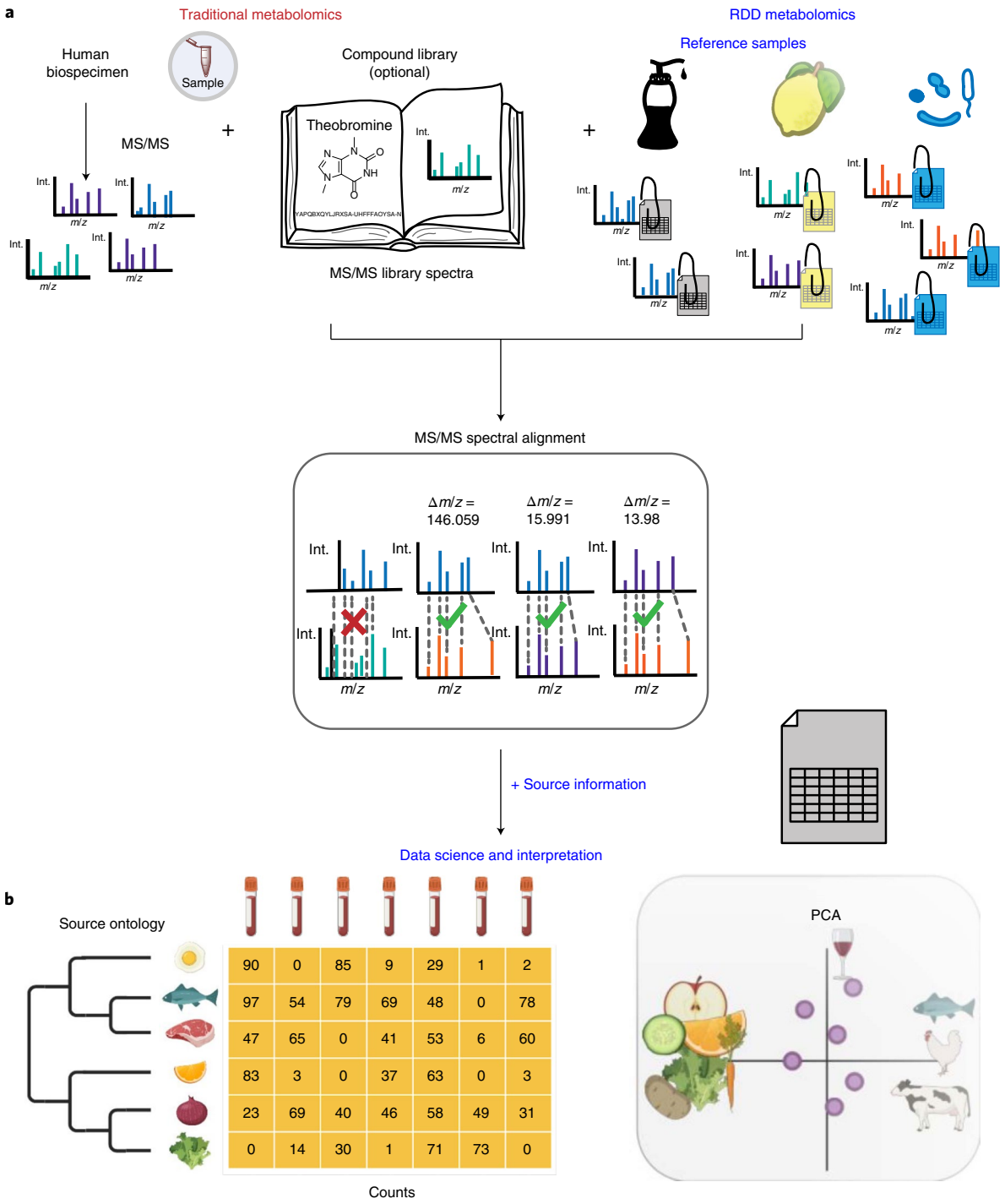


Fig. 1 | The concept of an RDD-based analysis workflow. a, Perform spectral alignment of the MS/MS-based untargeted metabolomics data from human biospecimens with data from reference samples that have controlled vocabularies for metadata. This can, optionally, be combined with MS/MS libraries. **b**, Link the spectral matches to the source information from the metadata from the reference samples. Create a data table of source ontology, human biospecimen and counts to enable data science and interpretation.

tain hundreds to thousands of MS/MS spectra, for metadata-based source annotation. The key differences are that the output reports contextualized information from source reference datasets. For successful RDD analysis, it is critical that the contextualized data are curated using controlled vocabularies or the results will not be amenable to downstream analysis. In the presented application for RDD, we investigated which food compositions could be recovered from data acquired from human biospecimens. Answering this question required a resource of reference food MS/MS source data and associated curated metadata. The source data includes MS/MS spectra of multiple ion forms of known and unknown molecules, isotopes, adducts, in-source fragments, and multimers[11,12]. The curated reference dataset can be matched in human biospecimens via direct matching of the MS/MS spectra or by molecular networking. Unlike static libraries, RDD analysis retains flexibility by enabling custom addition of files or metadata, and also gives

the user control on how the reference data is processed. We created a step-by-step tutorial for RDD analysis using Global Natural Products Social Molecular Networking (GNPS) (https://ccms-ucsd.github.io/GNPSDocumentation/tutorials/rdd/ and corresponding video tutorial https://www.youtube.com/watch?v=2-XsifrUY0Y)[13].

To exemplify RDD metabolomics, and because food is critical for health, we created a food metabolomics reference dataset. There is an unmet need to retrospectively and empirically read out food and beverage information from human metabolomics data, complementing current state-of-the-art mass spectrometry nutrition readout approaches targeting up to ~150–200 metabolites, food frequency and abundance questionnaires, diet records, 24-hr recalls, which can be self-monitored or assisted by a nutritional specialist[14,15]. The food reference dataset consists of untargeted metabolomics and detailed and structured metadata for ~3,500 foods (157 different food-specific metadata fields, Supplementary Table 1). It contains 107,968 unique MS/MS spectra merged from 1,907,765 spectra. The food source data can be easily expanded by creating and depositing additional datasets and metadata in GNPS/MassIVE.

For RDD, food source data is subjected to GNPS-based molecular networking[16,17] together with human metabolomics datasets (Fig. 2a). Using information on the controlled research diets of participants of a sleep and circadian study we assessed if RDD recovers food known to be consumed[18]. In this study, the participants were housed for four days, twice and were given a controlled diet, therefore we know if the results agreed with the known diet from that study (Fig. 2b). Of the 15 food categories, eleven represented direct matches to foods provided to the participants. Of those eleven matches, three matched to fermented versions of the non-fermented foods consumed such as fermented grapes instead of grapes, apple cider instead of apple, yogurt instead of milk, and four categories were not documented as consumed during the study, three of which could be explained. Evidence of caffeinated beverage consumption was observed only in two individuals—in the first 48 h in one volunteer and once in a second volunteer in the middle of the study—that there were few matches to caffeinated beverages is consistent with the elimination of caffeinated beverages in the controlled diet. Although not always written on the ingredient list of packages, rosemary is a common ingredient added to ground meat to slow oxidation and spoilage. The source of the matches to soda are unknown. This demonstrates that RDD can successfully obtain the correct diet information from untargeted metabolomics data but also be used to monitor diet adherence in controlled-diet studies.

We also tested mismatched food inventories by cross-matching US or Italian foods (different diets) and clinical cohorts. Crossover revealed that MS/MS spectral usage rates—the percentage of MS/MS spectra interpreted by the analysis—were 5–6% in reciprocal tests, versus 15–30% when the correct regional foods were used (Fig. 2c; $P=0.019$). These observations show that RDD analysis is selective on the basis of the foods that are consumed but also that it is important to continue to grow the food reference database as generic food databases have considerable value. Efforts, such as the Periodic Table of Food Initiative, and linking of Metabolights and Metabolomics workbench repositories with GNPS/MassIVE will aid the expansion of the food reference data.

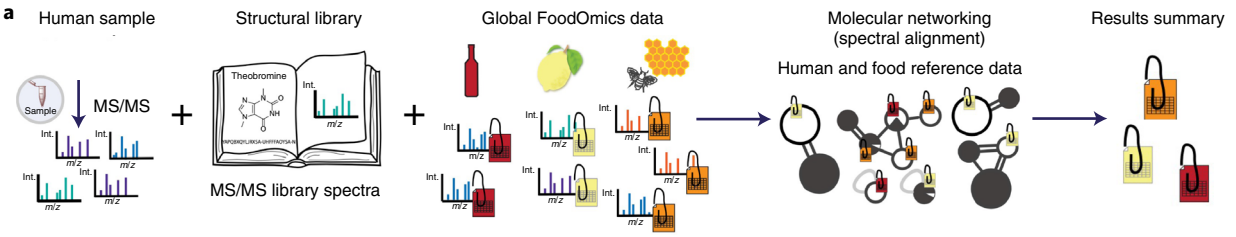
We next assessed if RDD analysis could recover a reference food spiked into human biospecimen extracts. We therefore analyzed mixtures of two human fecal samples or the NIST 1950 plasma reference extract with a tomato seedling extract in different proportions[19,20]. In all three biospecimens, the proportion of spectral matches relative to the tomato seedling extract increased linearly with the spiked-in proportion ($P=2.32\times10^{-31}$; Supplementary Fig. 1).

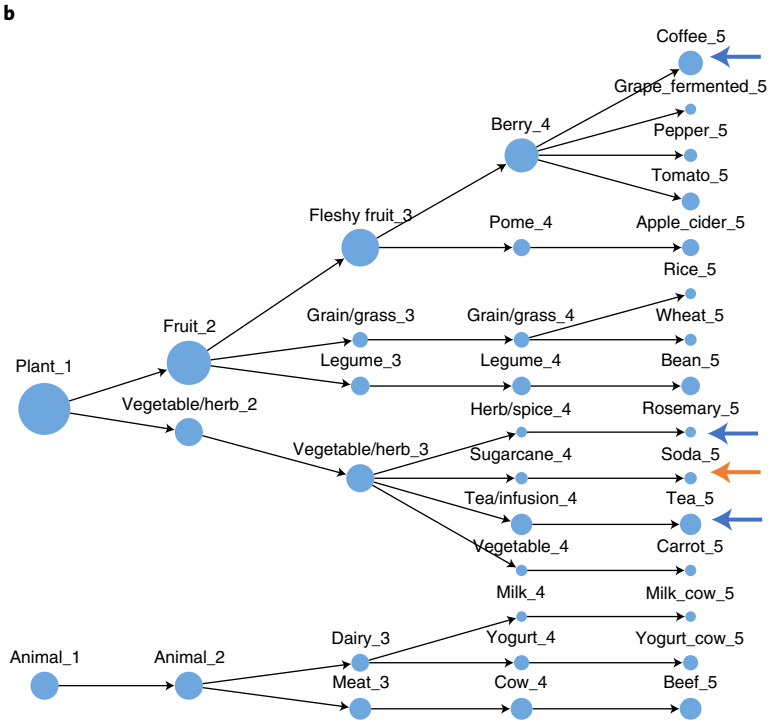
Because RDD analysis can be performed retrospectively, we co-analyzed the food reference dataset with 28 additional public human datasets (Supplementary Table 2, Supplementary Fig. 2). Of the MS/MS spectra, $10.1\pm4.4\%$ matched to spectral structural libraries. RDD increased MS/MS spectral usage 5.1 ± 3.3-fold over structural MS/MS library matches. With molecular networking, which can capture metabolized versions of molecules, spectral data usage increased 6.8 ± 3.5-fold. Inclusion of connected nodes, representing potential metabolism via molecular transformations, resulted in a total increase of $43.7\pm3.1\%$ (fecal; $P=6.9\times10^{-10}$), $51.2\pm6.9\%$ (plasma; $P=2.8\times10^{-6}$), and $58.0\pm4.2\%$ (other; $P=1.4\times10^{-6}$) of MS/MS spectra that can be leveraged as empirical readout of diet (Supplementary Fig. 2).

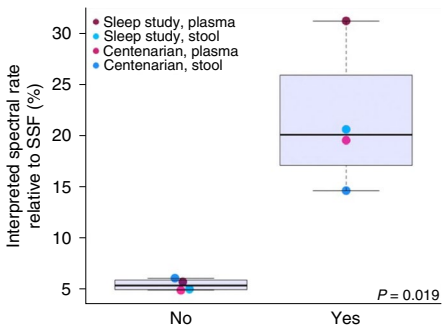
To validate the food consumption readouts obtained via RDD analysis from these 28 datasets, direct spectral library matches in the molecular networks created by the food-based RDD analyses (1% false discovery rate (FDR), and level 2/3 according to the metabolomics standards initiative[21,22]) were evaluated to verify whether they make sense in the context of food. An InChIKey is available for 4,586 of 5,455 spectral matches against the reference libraries, which yielded 1,492 unique structures upon consideration of planar structures. For 415 out of 1,492 planar structures that had lifestyle tags associated in GNPS[7,10], 'food consumption' was the most frequently reported tag (357 entries, 86%). Additionally, other matches are related to the food production chain, such as feed additives to promote animal growth that are tagged as 'drug', which include the antimicrobial agents monensin, enilconazole, kanamycin and other agricultural additives or environmental toxins (e.g. domoic acid)[23].

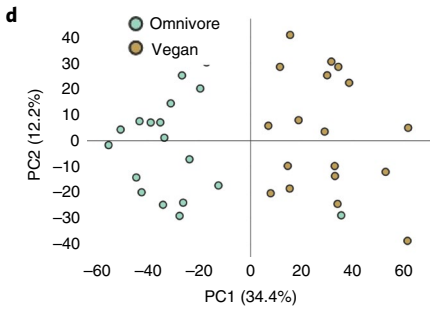
To assess if RDD can reveal dietary preferences, we analyzed a dataset of omnivores and vegans. Principal component analysis (PCA) of the spectral match relative proportions to reference foods revealed distinct patterns between dietary preferences (Fig. 2d). Omnivores had more MS/MS matches to dairy, meat, and seafood ($P=0.0021$, 2.2×10^{-10}, and 7.7×10^{-7}, respectively), while vegans had more MS/MS matches to legumes, fleshy fruit, and vegetables ($P=2.2\times10^{-10}$, 0.0096, and 0.029, respectively; Fig. 2e). Because many MS/MS spectra from foods may overlap, using only MS/MS spectra unique to each food can provide additional specificity (Fig. 2f). RDD analysis on an elderly population[24] found that

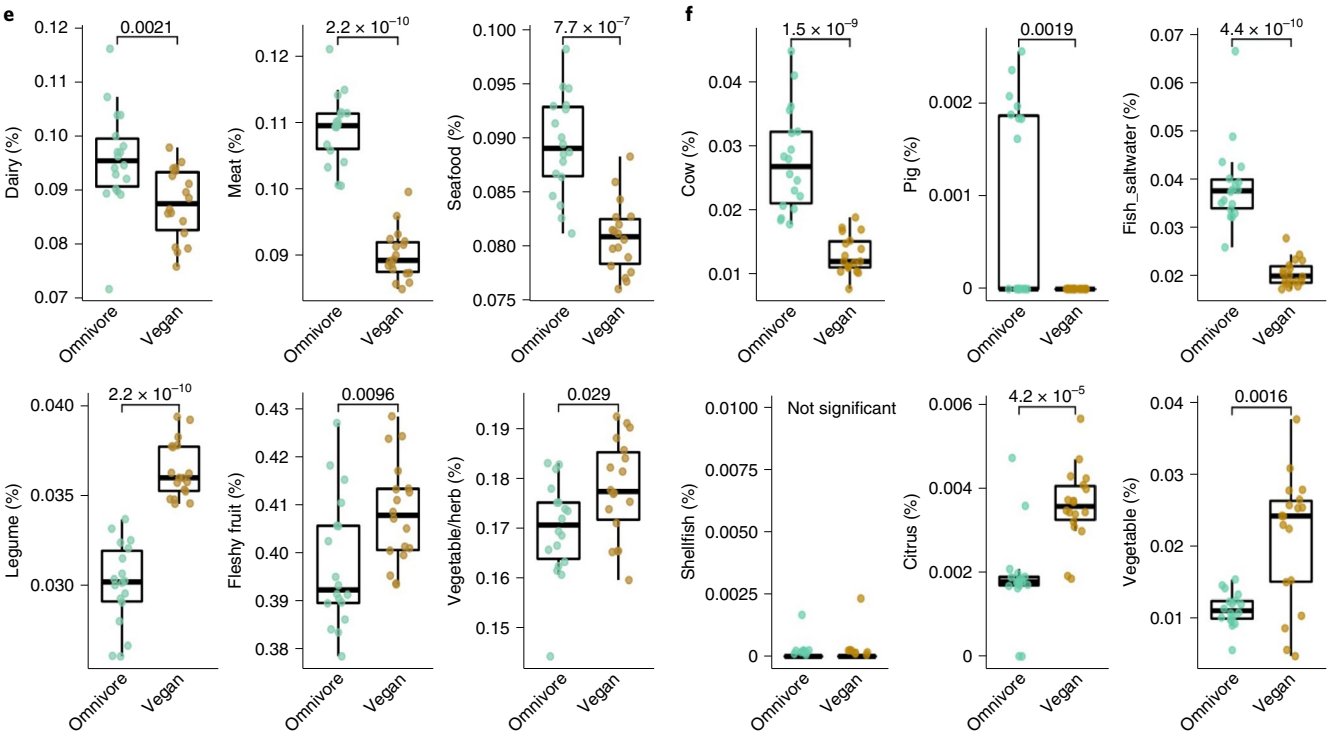
Fig. 2 | RDD with food reference data. a, Food RDD analysis schema. (int. = intensity) **b**, Food spectral counts (1% FDR[21]) observed in plasma from a sleep restriction and circadian misalignment study that controlled the diet of the participants ($n=371$ samples from 20 healthy adults)[18]. The size of node represents the relative number of spectral matches at each food level. Blue arrows indicate foods that could be explained although they were not provided in the study; orange arrow indicate source is not known. **c**, A crossover experiment between centenarian data from Italy and a sleep and circadian study from the US, for both fecal and plasma samples. Study-region-specific foods consumed by those individuals (yes) versus a different set of study-region-specific foods (no). One-way Welch's *t*-test, thick line is the mean, range within the box is the interquartile range (IQR) from the 25th to 75th quartile, whiskers indicate the minimum and maximum. **d**, PCA of food counts color coded by vegan (brown) versus omnivore data (green). **e**, Statistical analysis for the food counts at level 3 of the ontology, in relation to omnivore and vegan data (left six panels, dairy, meat, seafood, legume, fleshy fruit, vegetable, Wilcoxon test, $n=36$, 19 are vegan and 19 are omnivore). **f**, As in **e** but level 4 ontology using unique spectral counts (spectral usage is the percentage of MS/MS spectra used in the analysis. As they are unnamed ontologies as one would find in microorganism phylogeny in microbiome science (for example kingdom, genus, species) we have denoted these as layers (Right six panels, cow, pig, fish-saltwater, shellfish, citrus, vegetable, Supplementary Table 1). **e**,**f**, Boxes represent the IQR; the lower limit is the 25th percentile, the center line is the median, the upper limit is the 75th percentile; bars show the 75th percentile + 1.5 × IQR and the 25th percentile − 1.5 × IQR.


a
Human sample
Structural library
Global FoodOmics data
Molecular networking (spectral alignment)
Human and food reference data
Results summary
Sample
MS/MS
Theobromine
MS/MS library spectra


b
Plant_1
Fruit_2
Vegetable/herb_2
Fleshy fruit_3
Grain/grass_3
Legume_3
Vegetable/herb_3
Berry_4
Pome_4
Grain/grass_4
Legume_4
Herb/spice_4
Sugarcane_4
Tea/infusion_4
Vegetable_4
Milk_4
Dairy_3
Meat_3
Animal_1
Animal_2
Yogurt_4
Cow_4
Coffee_5
Grape_fermented_5
Pepper_5
Tomato_5
Apple_cider_5
Rice_5
Wheat_5
Bean_5
Rosemary_5
Soda_5
Tea_5
Carrot_5
Milk_cow_5
Yogurt_cow_5
Beef_5


c
Sleep study, plasma
Sleep study, stool
Centenarian, plasma
Centenarian, stool
Interpreted spectral rate relative to SSF (%)
P = 0.019
No
Yes


d
Omnivore
Vegan
PC2 (12.2%)
PC1 (34.4%)


e
Dairy (%)
0.0021
Meat (%)
2.2 × 10−10
Seafood (%)
7.7 × 10−7
Legume (%)
2.2 × 10−10
Fleshy fruit (%)
0.0096
Vegetable/herb (%)
0.029
f
Cow (%)
1.5 × 10−9
Pig (%)
0.0019
Fish_saltwater (%)
4.4 × 10−10
Shellfish (%)
Not significant
Citrus (%)
4.2 × 10−5
Vegetable (%)
0.0016
Omnivore
Vegan

individuals with lower diet diversity had more spectral matches to dairy, soda, and coffee, and this diet type was more prevalent in the group with Alzheimer's disease than those with normal cognition (Supplementary Fig. 3). This demonstrates that RDD analysis can be used to retrospectively stratify clinical studies onthe basis of empirical readout of diet composition for each sample.

RDD thus enables readout of dietary patterns (for example, vegan versus omnivore) and consumption of specific food items, and, more generally, can be used to match against any curated and ontology-aware reference database of sources, including environmental, or microbial sources. RDD metabolomics is currently unique to GNPS, as it requires highly scalable molecular networking and incorporation of detailed metadata. However, as other analysis ecosystems add molecular networking capabilities, or that make RDD compatible with other spectral alignment algorithms, it will become possible to use other resources for RDD metabolomics. As scalable molecular networking for GC–MS is also possible[25], specialized resources, such as BinBase[5,6], may eventually be leveraged for RDD analysis of specific applications or questions. To expand the scope of RDD metabolomics beyond food readout, well curated datasets of personal care products, medications (not just active ingredients but also formulations), microbial isolates, country of origin, biological sex, age, etc. might also be used as source reference data and requires careful curation with controlled vocabularies and structuring of metadata. Potential applications of RDD metabolomics include understanding diet and nutritional intake, exposure risks, medication use, consumption of illegal substances, environmental allergens, pollution studies, microbiome investigations, food ingredients/adulteration, forensics, and personal care product tracing to inform of potential exposures and health implications.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-022-01368-1.

Received: 24 June 2021; Accepted: 20 May 2022;
Published online: 07 July 2022

References

1. Knights, D. et al. Bayesian community-wide culture-independent microbial source tracking. *Nat. Methods* **8**, 8761–8763 (2011).
2. Ono, H. RefEx, a reference gene expression dataset as a web tool for the functional analysis of genes. *Scientific Data* **4**, 170105 (2017).
3. Bono, H. All of gene expression (AOE): an integrated index for public gene expression databases. *PLoS One* **15**, e0227076 (2020).
4. Turnbaugh, P. J. The human microbiome project. *Nature* **449**, 804–810 (2007).
5. Skogerson, K. et al. The volatile compound BinBase mass spectral database. *BMC Bioinf.* **12**, 321 (2011).
6. Lai, Z. et al. Identifying metabolites by integrating metabolome databases with mass spectrometry cheminformatics. *Nat. Methods* **15**, 53–56 (2018).
7. Bouslimani, A. et al. Lifestyle chemistries from phones for individual profiling. *Proc. Natl Acad. Sci.* **113**, E7645 (2016).
8. Haug, K. et al. MetaboLights: a resource evolving in response to the needs of its scientific community. *Nucleic Acids Res.* **48**, D440 (2020).
9. Damen, H. et al. Siscom—a new library search system for mass spectra. *Anal. Chim. Acta* **103**, 289–302 (1978).
10. Wang, M. et al. Mass spectrometry searches using MASST. *Nat. Biotechnology* **38**, 23–26 (2020).
11. Robin S., et al. *Nature Communications* **12**, 3832 (2021).
12. Li C., et al. Metabolite discovery through global annotation of untargeted metabolomics data. Preprint available at *bioRxiv* https://doi.org/10.1101/2021.01.06.425569 (2021).
13. Wang, M. et al. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
14. Barabási, A.-L. et al. The unmapped chemical complexity of our diet. *Nat. Food* **1**, 33–37 (2020).
15. Maruvada, P. et al. Perspective: Dietary Biomarkers of Intake and Exposure-Exploration with Omics Approaches. *Adv. Nutr.* **11**, 200–215 (2020).
16. Watrous, J. et al. Mass spectral molecular networking of living microbial colonies. *Proc. Natl Acad. Sci.* **109**, E1743–E1752 (2012).
17. Quinn, R. et al. Molecular networking as a drug discovery, drug metabolism, and precision medicine strategy. *Trends Pharmacol. Sci.* **38**, 143–154 (2017).
18. Sprecher, K. et al. Trait-like vulnerability of higher-order cognition and ability to maintain wakefulness during combined sleep restriction and circadian misalignment. *Sleep* **42**, zsz113 (2019).
19. Lungren, D. et al. Role of spectral counting in quantitative proteomics. *Expert Rev. Proteomics* **7**, 39–53 (2010).
20. Tripathi, T. et al. Chemically informed analyses of metabolomics mass spectrometry data with Qemistree. *Nat. Chem. Biol.* **17**, 146–151 (2021).
21. Scheubert, K. et al. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat. Commun.* **8**, 1494 (2017).
22. Sumner, L. et al. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative. *Metabolomics* **3**, 211–221 (2021).
23. West, K., et al., *NPJ Sci. Food* **6**, 22 (2022).
24. St. John-Williams, L. et al. Bile acids targeted metabolomics and medication classification data in the ADNI1 and ADNIGO/2 cohorts. *Scientific Data* **212**, 1 (2019).
25. Aksenov, A. et al. Auto-deconvolution and molecular networking of gas chromatography–mass spectrometry data. *Nat. Biotechnol.* **39**, 169–173 (2020).

[1]Collaborative Mass Spectrometry Innovation Center, University of California San Diego, La Jolla, CA, USA. [2]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California San Diego, La Jolla, CA, USA. [3]Beckman Center for Conservation Research, San Diego Zoo Wildlife Alliance, Escondido, CA, USA. [4]Center for Microbiome Innovation, Joan and Irwin Jacobs School of Engineering, University of California San Diego, La Jolla, CA, USA. [5]Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA. [6]Department of Pediatrics, School of Medicine, University of California San Diego, La Jolla, CA, USA. [7]Bioinformatics and Systems Biology Program, University of California San Diego, La Jolla, CA, USA. [8]Division of Rheumatology, Allergy & Immunology, Department of Medicine, University of California San Diego, La Jolla, CA, USA. [9]Division of Epidemiology and Community Health, School of Public Health, University of Minnesota, Minneapolis, MN, USA. [10]Department of Integrative Physiology, University of Colorado Boulder, Boulder, CO, USA. [11]Department of Population Health Sciences, University of Wisconsin-Madison, Madison, WI, USA. [12]Servicio Autónomo Centro Amazónico de Investigación y Control de Enfermedades Tropicales Simón Bolívar, Puerto Ayacucho, Amazonas, Venezuela. [13]Department of Pediatrics, Billings Clinic, Billings, MT, USA. [14]Department of Animal Science, Michigan State University, East Lansing, MI, USA. [15]Department of Pharmacology, Ribeirão Preto Medicinal School, Center of Research in Inflammatory Diseases, University of São Paulo, Ribeirão Preto, Sao Paolo, Brazil. [16]Multidisciplinary Health Institute, Federal University of Bahia, Vitória da Conquista, Bahia, Brazil. [17]NPPNS, Department of Biomolecular Sciences, School of Pharmaceutical Sciences of Ribeirão Preto, University of São Paulo, Ribeirão Preto, Sao Paolo, Brazil. [18]Department of Internal Medicine, Ribeirão Preto Medical School, Center of Research in Inflammatory Diseases, University of São Paulo, Ribeirão Preto, Sao Paolo, Brazil. [19]Department of Pharmacology, University of California San Diego, La Jolla, CA, USA. [20]Material Measurement Laboratory, National Institute of Standards and Technology, Gaithersburg, MD, USA. [21]Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Durham, Durham, NC, USA. [22]Department of Medicine, Duke University, Durham, NC, USA. [23]Duke Institute of Brain Sciences, Duke University, Durham, NC, USA. [24]Department of Neurosciences, University of California San Diego, La Jolla, CA, USA. [25]Division of Gastroenterology, Department of Medicine, University of California San Diego, La Jolla, CA, USA. [26]Department of Molecular Medicine, The Scripps Research Institute, La Jolla, CA, USA. [27]San Diego VA Healthcare System,

San Diego, CA, USA. [28]Division of Neonatology, Department of Pediatrics, Kapi'olani Medical Center for Women and Children, John A. Burns School of Medicine, Honolulu, Hawaii, USA. [29]Division of Neonatology, Perinatal Institute, Department of Pediatrics, Cincinnati Children's Hospital Medical Center, University of Cincinnati College of Medicine, Cincinnati, OH, USA. [30]Division of Pediatric Hospital Medicine, Department of Pediatrics, University of California San Diego, La Jolla, CA, USA. [31]Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences; Rutgers, The State University of New Jersey, New Brunswick, NJ, USA. [32]Department of Pediatrics, Washington University, St. Louis, MO, USA. [33]Department of Medicine, University of California San Diego, La Jolla, CA, USA. [34]Department of Computer Science and Engineering, University of California San Diego, La Jolla, CA, USA. [35]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA. [36]These authors contributed equally: Julia M. Gauglitz, Kiana A. West, Wout Bittremieux. ✉e-mail: rknight@ucsd.edu; pdorrestein@health.ucsd.edu

Methods

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 rheumatoid arthritis (MSV000084556; IRB 161474), late preterm (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), Malawi legume supplement (MSV000081486; IRB 201503171, Washington University Human Studies Committee), Rotarix vaccine response (MSV000084218; IRB PR-10060, 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 30248420.9.0000.5440, University of São Paulo, Brazil), IBD_biopsy (MSV000082220; IRB 120025), gout (MSV000084908; IRB 160768X), adult saliva (MSV000083049; IRB 150275, UCSD), legume supplementation (MSV000084663; IRB 201905103), NIST omnivore and vegan reference data (MSV000086989; de-identified NIST IRB MML-2019-035).

Global FoodOmics reference data. For the exemplary dataset used to highlight RDD metabolomics analysis we created and leveraged the 'Global FoodOmics' project (http://www.globalfoodomics.org) reference dataset. This dataset 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[26,27]. The majority of samples were photographed, and a subset were subjected to 16S ribosomal RNA 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 ribosomal RNA 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 Studies[28,29] (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 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 (2,399/3,579)). Primarily for the initial dataset, 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 −80 °C 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 UltraPure Distilled Water), as high ethanol concentrations are efficacious at preserving the sample for both DNA and metabolite analyses[30]. Samples were collected into 2-ml round bottom microcentrifuge tubes (Qiagen) and weighed before 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 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[31]. 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 −70 °C before 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); rheumatoid arthritis 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 owing to air pollution, a particular concern for 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 3,579 samples, which 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 1,133 fermented samples. This subcategorization of foods is made possible by the metadata collected on these samples, described below. 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 (Supplementary Table 1).

The foods, although primarily 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, and 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 the sample was collected, weight of the sample (pre-sample, post-sample, sample weight), day the sample was collected, and whether an image had been taken and renamed to match the sample number and archived in the image repository. The initial nine 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.

The above section describes the metadata for the food reference dataset, ideally one uses well-established controlled ontologies—if they allow one to answer the question the investigator cares about. For example, if one cares about the metabolic

changes in humans by latitude then the controlled metadata should have the latitude information. There are additional ontologies the user may want to use for answering different questions with RDD beyond the example provided here. In such cases, it is best to use an existing ontology, if available. There is an ontology lookup service at https://www.ebi.ac.uk/ols/index.

EMP[26], BIOM[32], REDU[33], and REDBIOM[34] are examples of systematic metadata capturing approaches that the authors have created previously. Proper metadata uses controlled vocabularies and is tedious and time consuming to collect in a systematic manner—usually taking more time than collecting the samples and data themselves—but is critical for the improved interpretation of the data.

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 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 six 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 on the basis of the known presence of common categories (corn, dairy*, egg*, fruit, fungi, fish*, shellfish*, meat, peanut*, seaweed, soy*, tree nut*, vegetable/herb, and wheat*, where asterisks designate 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; https://www.fda.gov/food/food-allergensgluten-free-guidance-documents-regulatory-information/food-allergen-labeling-and-consumer-protection-act-2004-falcpa), 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, owing 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 was then 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 versus rind), cheese_type (washed, blue etc), and cheese_texture (soft, semi-soft, semi-hard, and 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/). Wherever possible, we linked foods to food identifiers or created identifiers and categories that built upon the existing framework as defined by the US Department of Agriculture's Food and Nutrient Database for Dietary Studies 2011–2012 (FNDDS) food grouping scheme (https://www.ars.usda.gov/ARSUserFiles/80400530/pdf/fndds/fndds_2011_2012_doc.pdf). There are additional ontologies the user may want to use for answering different questions with RDD beyond what is captured here. In such cases, it is best to use an existing ontology, if available. There is an ontology look-up service at https://www.ebi.ac.uk/ols/index.

Metabolite extraction. The samples were suspended in 95% ethanol and homogenized in a tissue-lyser at 25 Hz for 5 min. Homogenized samples (in ethanol) were incubated for 40 min at −20 °C and centrifuged (Eppendorf centrifuge 5418) at 20,000 r.p.m. 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). 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). Resuspended extracts were then centrifuged for 15 min at 20,000 r.p.m. and 4 °C (Thermo SORVALL LEGEND RT) and transferred to a 96-well shallow-well plate, and diluted either 5× or 10× to avoid saturating the mass spectormetry detector.

Liquid chromatography–mass spectrometry. Food extracts were analyzed using an UltiMate 3000 ultra-high-performance liquid chromatography system (Thermo Scientific) equipped with a reverse phase C18 column, prepended with a guard cartridge (Kinetex, 100 × 2.1 mm, 1.7 µm particles size, 100 Å pore size; Phenomenex), at a column compartment temperature of 40 °C. Samples were chromatographically separated with a constant flow rate of 0.5 ml min^{-1} 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: H_2O + 0.1% formic acid; B: acetonitrile + 0.1% formic acid (LC–MS grade solvents, Fisher Chemical)).

The ultra-high-performance liquid chromatography system was coupled to a Maxis Q-TOF Impact II mass spectrometer (Bruker Daltonics) equipped with an electrospray ionization source. Mass spectra were acquired in positive ionization mode using data-dependent acquisition with a mass range of *m/z* 50–1,500. The instrument was externally calibrated two times per day to 1.0 p.p.m. mass accuracy using ESI-L Low Concentration Tuning Mix (Agilent Technologies). Hexakis (*m/z* 622.029509; (1H,1H,2H difluoroethoxy)phosphazene; Synquest Laboratories) was used for lock mass correction. MS/MS spectra were acquired for the top five ions in each MS1 spectrum, with active exclusion after two spectra (maintained for 30 s). Known contaminants as well as lock mass 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 GNPS (https://gnps.ucsd.edu), as described below.

FDR estimation. FDR estimation was calculated using Passatutto analysis workflow in GNPS[21,35]. FDR estimation was used to determine the cosine value required with a minimum of five matched peaks to achieve an FDR of 1%. See the Data Availability section for accession information.

Molecular networking using GNPS. In brief - molecular networking is accomplished by first merging all identical spectra of the study, structural reference libraries for annotations and food data using MS-Cluster[36]. Once merged, the merged spectra are aligned, taking in account the mass difference between the ions using a GNPS implementation of the modified cosine score. Throughout this process the metadata is tracked. Once the network has been created the resulting data table can then be used for downstream analysis. For the first report of the details of molecular networking see ref. [16], for the GNPS implementation of molecular networking see ref. [35], for a step-by-step instruction guide to molecular networking see ref. [37], for a review on use or interpretation of molecular networking see ref. [17].

Molecular networking analysis and library search were performed using GNPS classical molecular networking release_18[35]. 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)[38], HMDB[39], ReSpect[40], NIH natural product libraries[41], PNNL lipid library[42], Bruker/Sumner, FDA libraries, Gates Malaria library, EMBL library, as well as many other GNPS contributed libraries (https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp)[38] 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 (v.3.5.1)[43].

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 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.) (for example, number of stool nodes with a molecular ID/total number 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 (for example, number of nodes found in food and stool data/total number 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 of 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 – proportional 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. This was done for every analysis. 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 (v.3.8) using Pandas (v.0.25.3), NumPy (v.1.18.1), and floweaver (v.2.0.0a5)[44–46].

RDD metabolomics-based food counts does come with caveats to consider. First, because 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 (that is cholesterol produced by the human body versus consumed from meat) or that some molecules observed from animal sources such as vitamins (for example, pantothenate) or flavonoids are also observed in animals that consume them. 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 a specific relative proportion of 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.

Although we used all spectral matches in all figures except Fig. 2e,f where we used unique spectra only, care must be taken to not overinterpret the results, because some matches may get desired accuracy and precision only to level 1 of the ontology, but other matches may be precise and accurate all the way down to level 6. In other words, there are many more molecules that completely separate plants from animals (level 1) but are perhaps insufficient to readily separate out a red tomato from a yellow tomato (level 6). We show this directly in f. In f we explicitly use the unique MS/MS data only to get finer grained resolution. So instead of meat, we can now state (in proportions) who has more matches to pig meat or cow meat but that is only possible if there are unique spectra to that level. This is very similar to V4 amplification of 16S ribosomal RNA genes or related amplification methods in microbiome sequencing. In some cases, the data may allow for species identification, but most of the time only genus-level identification is possible. However, the V4 sequencing methodology is seeing extensive use to understand the microbiome. We also know that we are limited to the data of 3,600 foods for the comparisons, but this is only the beginning of the development of these approaches. In the next decade, we expect many new algorithms, more data availability (most in the metabolomics community still do not share their data publicly), and methods will be needed—especially as the reference database will get into the hundreds of thousands or even millions, but will continue to leverage reference data using concepts defined in this paper.

Recovery of spectra from a spiked-in reference sample. Two human fecal biospecimens and the NIST 1950 plasma reference were each mixed with increasing proportions of tomato seedling (*Solanum lycopersicum* plant) and analyzed using ultra high-performance liquid chromatography. This data was from a previous publication[20]. In brief, the samples were dissolved in 7/3 methanol/water and homogenized in a tissue lyser at 25 Hz for 5 min. The tubes were then centrifuged at 15,000 r.p.m. for 15 min and supernatant was collected. Extracts were then mixed in the following (biospecimen:seedling) ratios: 100:0, 75:25, 50:50, 25:75, and 0:100. The number of MS/MS matches between each sample and neat tomato seedling (reference sample, 0:100) were calculated. The significance of the linear relationship between seedling proportion and number of seedling spectral matches was tested using repeated measures correlation. The proportions of spectral matches between each sample and the reference sample, as well as each sample and non-plant food reference groups (at level 1 of the food ontology) were also calculated.

Diet information from the NIST omnivore and vegan reference data. Human whole stool was obtained from volunteer donors by the BioCollective. The samples consisted of whole stool from vegan and omnivore donors (four 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 package in R. Counts were aggregated for specific food categories (dairy, meat, seafood, legume, fleshy fruit, and 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 patients with Alzheimer's disease. As described above, a feature table was generated on the basis of 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 2,123 samples, representing multiple different biofluids and tissues (Supplementary Table 1).

Data were collected for the following studies using a quadropole time-of-flight 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)[47], 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 rheumatoid arthritis (MSV000084556) and Alzheimer's disease (MSV000085256) were collected with similar methods on a Q-exactive Orbitrap mass spectrometer (Thermo Scientific). The Alzheimer's 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. 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, on the basis of 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 stipulations 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 body weight; 4) admission electrocardiogram without changes in rhythm and with QT interval <450 ms; 5) normal serum levels of Ca^{2+} 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) were pregnant; 3) were unable to understand the information contained in the Free and Informed Consent Form.

Sample preparation: for the COVID-19 plasma samples, aliquots of 20 µl were transferred to Eppendorf tubes and 120 µl cold extracting solution, MeOH:MeCN (1:1, vol/vol) was added. After orbital shaking for 1 min (Gehaka AV-2 Shaker), the samples were left at −20 °C for 30 min and then centrifuged for 10 min at 20,000*g* at 4 °C (Centrifuge Boeco Germany M-240R). 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 30 °C (Analitica, Christ RVC2-18). The residues were resuspended in 80 µl H_2O and centrifuged (10 min, 5,000*g*, 4 °C), an aliquot of 5 µl was injected.

For mass spectrometry data collection of plasma sample, extracts were chromatographically separated with an HPLC (Shimadzu), coupled with a micrOTOF-Q II mass spectrometer (Bruker Daltonics) equipped with an ESI source and a quadrupole-time of flight analyzer (Bruker Daltonics Inc.). For chromatographic analyses, we employed a Kinetex C18 column (1.7 µm, 100 × 2.1 mm) (Phenomenex) kept at 40 °C, with a flow rate of 0.3 ml min^{-1}. 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 (vol/vol) and phase B is acetonitrile 0.1% formic acid (vol/vol) (LC–MS grade solvents). The MS data were acquired in positive mode using an MS range of *m/z* 50–1,500. The equipment was calibrated with trifluoroacetic acid every day, and internally during each run. The MS parameters were established as follows: end plate offset, 450 V; capillary voltage, 3,500 V; nebulizer gas pressure, 4.0 Bar; dry gas flow, 9 l min^{-1}; dry temperature, 220 °C.

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 s. A fragmentation exclusion list was set to exclude known contaminants and infused lock mass compounds: *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; and 959.80–962.00. 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) in 50% MeOH (LC–MS grade solvents) was injected every five samples. All MS data were analyzed with Bruker Compass DataAnalysis 4.3 software (Bruker Daltonics).

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 (trifluoroacetic acid), converted to .mzXML files using MSConvert from the ProteoWizard software 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.

Resources to get started on your own dataset. There is a recorded introduction workshop that was given as part of the Shaping the Microbiome Through Nutrition UCSD-Nature Publishing conference. https://ccms-ucsd.github.io/GNPSDocumentation/workshops/. For a step-by-step guide and video see https://ccms-ucsd.github.io/GNPSDocumentation/tutorials/rdd/ and corresponding video tutorial https://www.youtube.com/watch?v=2-XsifrUY0Y.

Reporting summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The following files are available in addition to the Global FoodOmics mzXML files on https://massive.ucsd.edu under MSV000084900: metadata as a.txt; an image repository with between one and six 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 Fig. 1). A metadata dictionary can also be accessed here: https://docs.google.com/spreadsheets/d/1Ebn-TgMWEkd_7KOw9TCRvHGPsE7dGjVCr7dg28pwbmM/edit#gid=727944641. The accessions numbers to the raw metabolomics data files available via Supplementary Table 2. The GNPS-based molecular networking analyses jobs 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 rheumatoid arthritis (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); Alzheimer's disease serum (MSV000086270, https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=570aacf2244948c7afa590631de5d345); omnivore versus vegan (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); tomato seedling (MSV000083353, https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=3b6020d7034045c39969631894ae4c22); food only (MSV000084900, https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d5adba7f67cc402396e9ba7cd85ce52b). Networking parameters were set on the basis of 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 rheumatoid arthritis, 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; Alzheimer's disease, 6581031643254259 81c097cecba840b0; Alzheimer's disease serum, 67516099b37647f2a9c91f890366bef3; omnivore versus vegan, ba974d08cab04f77aaacdb7828baada6; gout, a478f419ae824378aa02e5e1b310cad2; adult saliva, 32980f95dbd5437aaa9e15d05c7246bb; LP infant, 8bfbdc1bf38c418fb223306cd42af897; LP infant, 3e414e13a4394bb78c07f7ca7f4d1be3; legume supplementation, 2ca007303b9c4bb3820f392b996eba27; COVID-19 Brazil, d16eb32276c84bdb9c35c5872e97a986; Tomato seedling, f1c9cd79e0e94c66a367b6816b149750.

Code availability

The code generated during this study is available at https://github.com/DorresteinLaboratory/GlobalFoodomics.

References

26. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).
27. McDonald, D. et al. American Gut: an open platform for citizen science microbiome research. *mSystems* **3**, e00018-31 (2018).
28. Sicherer, S. H. & Sampson, H. A. Food allergy: A review and update on epidemiology, pathogenesis, diagnosis, prevention, and management. *J. Allergy Clin. Immunol.* **117**, S470–S475 (2006).
29. Martin, C. L., et al. *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, 2012).
30. Song, S. J. et al. Preservation methods differ in fecal microbiome stability,affecting suitability for field studies. *mSystems* **1**, e00021-16 (2016).
31. Sprecher, K. J. et al. Trait-like vulnerability of higher-order cognition and ability to maintain wakefulness during combined sleep restriction and circadian misalignment. *Sleep* **42**, zsz113 (2019).
32. McDonald, D. et al. The Biological Observation Matrix (BIOM) format or: how I learned to stop worrying and love the ome-ome. *Gigascience.* **1**, 7 (2012).
33. Jarmusch, A. K. et al. ReDU: a framework to find and reanalyze public mass spectrometry data. *Nat. Methods* **17**, 901–904 (2020).
34. McDonald, D. et al. redbiom: a rapid sample discovery and feature characterization system. *mSystems* **4**, e00215-19 (2019).
35. Wang, M. et al. Sharing and community curation of mass spectrometry data with global natural products social molecular networking. *Nat. Biotechnol.* **34**, 828–837 (2016).
36. Frank, A. M. et al. Spectral archives: extending spectral libraries to analyze both identified and unidentified spectra. *Nat. Methods* **8**, 587–591 (2011).
37. Aron, A. T. et al. Reproducible molecular networking of untargeted mass spectrometry data using GNPS. *Nat. Protoc.* **15**, 1954–1991 (2020).
38. Horai, H. et al. Massbank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
39. Wishart, D. S. et al. HMDB 4.0: the human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
40. Sawada, Y. et al. RIKEN tandem mass spectral database (ReSpect) for phytochemicals: a plant-specific MS/MS-based data resource and database. *Phytochemistry* **82**, 38–45 (2012).
41. Huang, R. et al. The NCATS pharmaceutical collection: a 10-year update. *Drug Discov.* **24**, 2341–2349 (2019).
42. Kyle, J. E. et al. LIQUID: an-open source software for identifying lipids in LC–MS/MS-based lipidomics data. *Bioinformatics.* **33**, 1744–1746 (2017).
43. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
44. McKinney, W. Data Structures for Statistical Computing in Python. *In Proc. 9th Python in Science Conference* (Eds. van der Walt, S. & Millman, J.) 56–61 (SciPy, 2010).
45. van der Walt, S., Colbert, S. C. & Varoquaux, G. The NumPy array: a structure for efficient numerical computation. *Comput. Sci. Eng.* **13**, 22–30 (2011).
46. Lupton, R. C. & Allwood, J. M. Hybrid Sankey diagrams: visual analysis of multidimensional data for understanding resource use. *Resour. Conserv. Recycl.* **124**, 141–151 (2017).
47. Taylor, B. C. et al. Consumption of fermented foods is associated with systematic differences in the gut microbiome and metabolome. *mSystems* **5**, e00901-19 (2020).

Acknowledgements

Funding sources: we thank the Crohn's & Colitis foundation #675191, U19 AG063744 01, R01AG061066, 1 DP1 AT010885, P30 DK120515, Office of Naval Research MURI grant N00014-15-1-2809 and NIH/NCATS Colorado CTSA Grant UL1TR002535, the Emch Fund and C&D Fund. This work was also supported in part by the Chancellor's Initiative in the Microbiome and Microbial Sciences and by Illumina through reagent donation and by Danone Nutricia Research in partnership with the Center for Microbiome Innovation at UCSD. We would like to thank E. Sayyari, D. S. Nguyen, E. Wolfe and K. Sanders for sample processing, and J. DeReus for data handling, processing, and maintaining the computational infrastructure. J.P.S. was supported by SD IRACDA (5K12GM068524-17), and in part by USDA-NIFA (2019-67013-29137) and the Einstein Institute GOLD project (R01MD011389). R.C. and M.G. were supported by the Krupp Endowed Fund; R.C. was also supported by a UCSD Rheumatic Diseases Research Training Grant from the NIH/NIAMS (T32AR064194). VA Research Service, NIH/NIAMS AR060772 and AR075990 to R.T., R.H.M. was supported through a UCSD training grant from the NIH/NIDDK Gastroenterology Training Program (T32 DK007202). The Brazilian National Council for Scientific and Technological Development (CNPq)-Brazil (245954/2012) to M.F.O. and FAPESP (2014/50265-3) to N.P.L. D.W. was supported by NIH/NHLBI Training Grant (NIH T32 HL149646). K.S. was supported by a PROMOS fund (DAAD). W.B. is a postdoctoral researcher of the Research Foundation–Flanders (FWO). R.J.D. was supported by NIH DP2 AT010401-01. We thank R. da Silva for his feedback and early bioinformatics analysis for the Global FoodOmics project. We further acknowledge all the individuals that contributed samples as well as companies and organizations that have donated samples: D. Vargas, Townshend's Tea Company, BDK Kombucha, Oregonian Tonic, Squirrel & Crow, Venissimo cheese, Fermenter's Club San Diego, Good Neighbor Gardens, Sprouts Farmers Market, Ralphs, Whole Foods, Julian Ciderworks and San Diego Zoo and Safari Park. Specifically thank you to A. Durant for coordinating sampling at Fermentation Festivals and the wonderful staff at San Diego Zoo Wildlife Alliance for coordinating and helping with sample collection: M. Gaffney, E. Galindo, K. Kerr, A. Fidgett, J. Stuart, D. Tanciatco, and L. Pospychala. NIST would like to acknowledge The Institute for the Advancement of Food and Nutrition Sciences (IAFNS) microbiome committee for providing support for the development of standardized fecal materials. Funding for the ADMC (Alzheimer's Disease Metabolomics Consortium, led by Dr R.K.-D. at Duke University) was provided by the National Institute on Aging grants 1U01AG061359-01 and R01AG046171, a component of the Accelerating Medicines Partnership for AD (AMP-AD) Target Discovery and Preclinical Validation Project (https://www.nia.nih.gov/research/dn/ampad-target-discovery-and-preclinical-validation-project) and the National Institute on Aging grant RF1 AG0151550, a component of the M2OVE-AD Consortium (Molecular Mechanisms of the Vascular Etiology of AD Consortium, https://www.nia.nih.gov/news/decoding-molecular-tiesbetween-vascular-disease-and-alzheimer). Additional support was provided by the following NIA grants: (1RF1AG058942-01 and 3U01 AG024904-09S4). Data collection and sharing for the ADNI was supported by National Institutes of Health Grant U01 AG024904. ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd; Janssen Alzheimer Immunotherapy Research & Development, LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. UCSD Academic Senate Research/Bridge Grant. Eunice Kennedy Shriver National Institute of Child Health and Human Development K12-HD000850.

Author contributions

P.C.D., R.K.D., R.J.D., and J.M.G. conceptualized the idea. M.J.M., M.B., M.P., F.D.O., K.C.W., C.M.A., E.B., K.S., P.C.D., R.J.D., R.K.D., N.C.S., A.D.S., K.D., G.A., D.M.D., N.P.L., M.B., and J.M.G. collected FoodOmics samples and performed metadata curation. M.J.M, M.P., F.D.O., F.V., C.M.A., E.B., N.C.S., and J.M.G. performed FoodOmics sample processing and MS data acquisition. A.J.J., P.B.F., E.D., Q.Z., D.N., D.M., J.P.S., and J.M.G. curated Global FoodOmics metadata to match FNDDS. K.E.R., J.B.W., B.S.B., B.J.B., R.C.,

M.G.D.B., M.M.D., E.O.E., D.G., L.H., J.H.K., M.M., C.M., R.K., K.E.S., D.V.R., T.I.K., C.W., K.P.W.J., M.F.O., R.H.M., D.W., R.T., J.G.A., P.S.D., M.G., D.J.G., A.K.J., B.J.B., R.M.S., K.C.W., A.D.S., F.V., N.P.L., P.K.P., S.M.D.S., S.L.S., C.M.J., N.J.L., K.A.L., S.A.J., R.K.D. and J.M.G. provided samples, comparative dataset, and/or detailed metadata. L.M.M.M., T.M.C. performed COVID-19 patient and/or food sample preparation and analysis. P.L.J. was the physician responsible for the COVID-19 patients. R.D.R.O was the physician responsible for collecting the plasma from COVID-19 patients. F.P.V. was responsible for tabulation of COVID-19 patient data. M.P., J.M.G., T.S., M.G.D.B., L.D.R.G., G.H. prepared samples for food. M.W. supported GNPS computational infrastructure used in the study. C.L.W., W.B., A.K.J., K.A.W., E.S., A.T., N.P.L. and J.M.G. analyzed MS data. C.L.W., W.B., A.K.J., K.A.W., C.M., and J.M.G. generated figures. P.C.D., R.K., R.J.D., A.D.S., and J.M.G. supervised the work. P.C.D., R.K., C.L.W., K.A.W., W.B., and J.M.G. wrote the paper. All authors have contributed feedback and edits to the manuscript.

Competing interests

B.S.B. has a research grant from Prometheus Biosciences and has received consulting fees from Pfizer. P.C.D. is on the scientific advisory board of Sirenas, Cybele Microbiome, Galileo, and founder and scientific advisor of Ometa Labs LLC and Enveda (with approval by UC San Diego). J.H.K. is a consultant for Medela and on the Board for Innara Health; he owns shares in Astarte Medical and Nicolette. M.G. has research grants from Pfizer and Novartis. P.S.D. has received research support and/or consulting from Takeda, Pfizer, Abbvie, Janssen, Prometheus, Buhlmann, Polymedco. R.J.D. is a consultant for and owns shares in Impossible Foods Inc., and is on the Scientific Advisory Panel of Boost Biomes. A.J.J. has received consulting fees from Abbott Nutrition and Corebiome. D.G. is a consultant for Biogen, Fujirebio, vTv Therapeutics, Esai and Amprion and serves on a DSMB for Cognition Therapeutics. K.P.W. reports during the conduct of the study receiving research support from SomaLogic, Inc., consulting fees from or served as a paid member of scientific advisory boards for the Sleep Disorders Research Advisory Board–National Heart, Lung and Blood Institute, CurAegis Technologies, Philips, Inc., Circadian Therapeutics, Ltd. and Circadian Biotherapies Ltd. R.T. received a research grant from AstraZeneca Consulting, SOBI, Selecta, Horizon, Allena, AstraZeneca. A.D.S. and R.K. are directors at the Center for Microbiome Innovation at UC San Diego, which receives industry research funding for multiple microbiome initiatives, but no industry funding was provided for this project. M.W. is a co-founder of Ometa Labs LLC. K.D. is an inventor on a series of patents on the use of metabolomics for the diagnosis and treatment of central nervous system diseases and holds equity in Metabolon Inc., Chymia LLC and PsyProtix. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01368-1.

Correspondence and requests for materials should be addressed to Rob Knight or Pieter C. Dorrestein.

Peer review information *Nature Biotechnology* thanks Elaine Holmes and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.