## Title

Enhancing untargeted metabolomics using metadata-based source annotation

**Permalink**
https://escholarship.org/uc/item/79t6410z

**Journal**
Nature Biotechnology, 40(12)

**ISSN**
1087-0156

**Authors**
Gauglitz, Julia M
West, Kiana A
Bittremieux, Wout
et al.

**Publication Date**
2022-12-01

**DOI**
10.1038/s41587-022-01368-1

Peer reviewed

Published in final edited form as:

*Nat Biotechnol*. 2022 December ; 40(12): 1774–1779. doi:10.1038/s41587-022-01368-1.

# Enhancing untargeted metabolomics using metadata-based source annotation

*A full list of authors and affiliations appears at the end of the article.*

## Abstract

Human untargeted metabolomics studies annotate only ~10% of molecular features. We introduce reference data–driven analysis to match metabolomics 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.

## Editorial summary:

[#]co-first author.

Author Contributions:
PCD, RK, RJD, and JMG conceptualized the idea.
MP, FDO, KCW, CMA, EB, KS, PCD, RJD, RK, NCS, ADS, KD, GA, DM, NPL, and JMG collected FoodOmics samples and performed metadata curation.
MP, FDO, FV, CMA, EB, NCS, and JMG performed FoodOmics sample processing and MS data acquisition.
AJJ, PBF, ED, QZ, DN, DM, JPS, and JMG curated Global FoodOmics metadata to match FNDDS.
JBW, BSB, BJB, RC, MGDB, MD, EOE, DG, LH, JHK, MM, CM, RK, KES, DVR, TIK, CW, KPW, MFO, RHM, DW, RT, JGA, PD, MG, DG, AKJ, BJB, RMS, KCW, ADS, FV, NPL, PKP, SMS, SLS, CMJ, NJL, KAL, SAJ, RKD and JMG provided samples, comparative dataset, and/or detailed metadata.
LMMM, TMC performed COVID-19 patient and/or food sample preparation and analysis.
PLJ was the physician responsible for the COVID-19 patients.
RDRO was the physician responsible for collecting the plasma from COVID-19 patients.
FPV was responsible for tabulation of COVID-19 patient data.
MP, JMG, TS, MB, LDG, GH prepared samples for food.
MW supported GNPS computational infrastructure used in the study.
CLW, WB, AKJ, KAW, ES, AT, NPL and JMG analyzed MS data.
CLW, WB, AKJ, KAW, CM, and JMG generated figures.
PCD, RK, RJD, ADS, and JMG supervised the work.
PCD, RK, CLW, KAW, WB, and JMG wrote the paper.
All authors have contributed feedback and edits to the manuscript.

Declaration of Interests:
BSB has a research grant from Prometheus Biosciences and has received consulting fees from Pfizer. PCD 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). JHK is a consultant for Medela and on the Board for Innara Health; he owns shares in Astarte Medical and Nicolette. MG has research grants from Pfizer and Novartis. PSD has received research support and/or consulting from Takeda, Pfizer, Abbvie, Janssen, Prometheus, Buhlmann, Polymedco. RJD is a consultant for and owns shares in Impossible Foods Inc, and is on the Scientific Advisory Panel of Boost Biomes. AJJ has received consulting fees from Abbott Nutrition and Corebiome. DG is a consultant for Biogen, Fujirebio, vTv Therapeutics, Esai and Amprion and serves on a DSMB for Cognition Therapeutics. KPW 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. RT: Research grant: Astra-Zeneca Consulting: SOBI, Selecta, Horizon, Allena, Astra-Zeneca. ADS and RK 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. MW is a co-founder of Ometa Labs LLC. Dr. Kaddurah-Daouk in an inventor on a series of patents on use of metabolomics for the diagnosis and treatment of CNS diseases and holds equity in Metabolon Inc., Chymia LLC and PsyProtix. The remaining authors declare no competing interests.

Disclaimer: Certain commercial equipment, instruments, software or materials are identified in this document. Such identification does not imply recommendation or endorsement by the National Institute of Standards and Technology, nor does it imply that the products identified are necessarily the best available for the purpose.

Metabolomics is improved by using a reference library of both known and unknown molecules.

Complex sequence data from metagenomic 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 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 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.[18,19] In addition, as we have previously demonstrated, using structural annotations, the source can be determined by literature mining.[20] However, due to 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 data set level. We therefore introduce the RDD approach for metabolomics (Figure 1), followed by a use case demonstrating empirical food readouts from untargeted human data (Figure 2).

Untargeted MS/MS-based metabolomics experiments involve searching MS/MS structural libraries since the late 1970’s [5,6], or, more recently, for investigating the distribution of a MS/MS spectrum across public untargeted data.[21] 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 contain 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.[7,8] 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 GNPS (https://ccms-ucsd.github.io/GNPSDocumentation/tutorials/rdd/ and corresponding video tutorial https://www.youtube.com/watch?v=2-XsifrUY0Y).[9]

To exemplify RDD metabolomics, and because food is critical for health, we created a food metabolomics reference data set. 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. [10,11] The food reference data set consists of untargeted metabolomics and detailed and structured metadata for ~3500 foods (157 different food-specific metadata fields, Table S1). 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 data sets and metadata in GNPS/MassIVE.

For RDD, food source data is subjected to GNPS based molecular networking[14,15] together with human metabolomics datasets (Figure 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[12]. In this study, the participants were housed for four weeks and were given a controlled diet, therefore we know if the results agreed with the known diet from that study (Figure 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 48hrs 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 (Figure 2c, p=0.019). These observations show that RDD analysis is selective based on 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.[22] In all three biospecimens, the proportion of spectral matches relative to the tomato seedling extract increased linearly with the spiked-in proportion ($p=2.32E^{-31}$, SI Figure 1).

Because RDD analysis can be performed retrospectively, we co-analyzed the food reference dataset with 28 additional public human datasets (Table S2, SI Figure 2). 10.1±4.4% of MS/MS spectra 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±3.1% (fecal; $P$=6.9e-10), 51.2±6.9% (plasma; $P$=2.8e-06), and 58.0±4.2% (other; $P$=1.4e-06) of MS/MS spectra that can be leveraged as empirical readout of diet (SI Figure 2).

To validate the food consumption read-outs obtained via RDD analysis from these 28 datasets, direct spectral library matches in the molecular networks created by the food-based RDD analyses (1% FDR, and level 2/3 according to the metabolomics standards initiative [13, 17]) 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 1492 planar structures that had lifestyle tags associated in GNPS [20,21], “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”, such as the antimicrobial agents monensin, enilconazole, kanamycin and other agricultural additives or environmental toxins such as domoic acid.[25]

To assess if RDD can reveal dietary preferences, we analyzed a data set of omnivores and vegans. Principal component analysis (PCA) of the spectral match relative proportions to reference foods revealed distinct patterns between dietary preferences (Figure 2d). Omnivores had more MS/MS matches to dairy, meat, and seafood ($P$=0.0021, 2.2e-10, and 7.7e-7 respectively), while vegans had more MS/MS matches to legumes, fleshy fruit, and vegetables ($P$=2.2e-10, 0.0096, and 0.029, respectively, Figure 2e). Because many MS/MS spectra from foods may overlap, using only MS/MS spectra unique to each food can provide additional specificity (Figure 2f). RDD analysis on an elderly population[16] found that individuals with lower diet diversity had more spectral matches to dairy, soda, and coffee, and this diet type was more prevalent in the Alzheimer’s Disease group than those with normal cognition (SI Figure 3). This demonstrates that RDD analysis can be used to retrospectively stratify clinical studies based on empirical readout of diet composition for each sample.

RDD thus enables readout of dietary patterns (e.g. 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[24], specialized resources, such as BinBase[18,19], 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.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Authors

Julia M. Gauglitz[1,2,#], Kiana A. West[1,2,#], Wout Bittremieux[1,2,#], 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[27], David J. Gonzalez[1,2,4,19], Robert Terkeltaub[8,34], Megan M. Doty[6,27], Jae H. Kim[28], Kyung E. Rhee[6], Julia Beauchamp-Walters[29], Kenneth P. Wright Jr[10], Maria Gloria Dominguez-Bello[29], Mark Manary[30], Michelli F. Oliveira[31], Brigid S. Boland[21], Norberto Peporine Lopes[17], Monica Guma[8], Austin D. Swafford[4], Rachel J. Dutton[5], Rob Knight[4,6,32,33], Pieter C. Dorrestein[1,2,4,6]

## Affiliations

[1] Collaborative Mass Spectrometry Innovation Center; University of California San Diego; La Jolla, CA 92093; USA

[2] Skaggs School of Pharmacy and Pharmaceutical Sciences; University of California San Diego; La Jolla CA 92093; USA

[3] Beckman Center for Conservation Research; San Diego Zoo Wildlife Alliance; Escondido, CA 92027; USA

[4] Center for Microbiome Innovation, Joan and Irwin Jacobs School of Engineering; University of California San Diego; La Jolla, CA 92093; USA

[5] Division of Biological Sciences; University of California San Diego; La Jolla, CA 92093; USA

[6] Department of Pediatrics, School of Medicine; University of California San Diego; La Jolla, CA 92093; USA

[7] Bioinformatics and Systems Biology Program; University of California San Diego; La Jolla, CA 92093; USA

[8] Division of Rheumatology, Allergy & Immunology, Department of Medicine; University of California San Diego; La Jolla, CA 92093; USA

[9] Division of Epidemiology and Community Health, School of Public Health; University of Minnesota; Minneapolis, MN 55455; USA

[10] Department of Integrative Physiology; University of Colorado Boulder; Boulder, CO 80309; USA

[11] Department of Population Health Sciences; University of Wisconsin-Madison; Madison, WI 53726; USA

[12] Servicio Autónomo Centro Amazónico de Investigación y Control de Enfermedades Tropicales Simón Bolívar; Puerto Ayacucho 7101, Amazonas; Venezuela

[13] Department of Pediatrics; Billings Clinic; Billings, Montana 59101; USA

[14] Department of Animal Science; Michigan State University; East Lansing, MI 48824; USA

[15] Department of Pharmacology, Ribeirão Preto Medicinal School, Center of Research in Inflammatory Diseases; University of São Paulo; Ribeirão Preto, CEP 14049-900 - SP; Brazil

[16] Multidisciplinary Health Institute; Federal University of Bahia; 45029094, Vitória da Conquista - BA; Brazil

[17] NPPNS, Department of Biomolecular Sciences, School of Pharmaceutical Sciences of Ribeirão Preto; University of São Paulo; Ribeirão Preto. CEP 14040-903 - SP; Brazil

[18] Department of Internal Medicine, Ribeirão Preto Medical School, Center of Research in Inflammatory Diseases; University of São Paulo; Ribeirão Preto, CEP 14049-900 - SP; Brazil

[19] Department of Pharmacology; University of California San Diego; La Jolla, CA 92093; USA

[20] Material Measurement Laboratory; National Institute of Standards and Technology, Gaithersburg, MD 20899; USA

[21] Department of Psychiatry and Behavioral Sciences, Duke University School of Medicine, Durham, Durham, NC, 27710, USA.

[22] Department of Medicine, Duke University, Durham, NC, 27710, USA.

[23] Duke Institute of Brain Sciences, Duke University, Durham, NC, 27710, USA.

[24] Department of Neurosciences; University of California San Diego; La Jolla, CA 92093; USA

[25] Division of Gastroenterology, Department of Medicine; University of California San Diego; La Jolla, CA 92093; USA

[26] Department of Molecular Medicine; The Scripps Research Institute; La Jolla, CA 92037; USA

[27] Division of Neonatology, Department of Pediatrics, Kapi'olani Medical Center for Women and Children; John A. Burns School of Medicine; Honolulu, Hawaii 96813; USA

[28] Division of Neonatology, Perinatal Institute, Department of Pediatrics, Cincinnati Children's Hospital Medical Center; University of Cincinnati College of Medicine; Cincinnati, Ohio 45229; USA

[29] Division of Pediatric Hospital Medicine, Department of Pediatrics,; University of California San Diego; La Jolla, CA 92093; USA

[30] Department of Biochemistry and Microbiology, School of Environmental and Biological Sciences; Rutgers, The State University of New Jersey; New Brunswick, NJ 08901; USA

[31] Department of Pediatrics; Washington University; St. Louis, MO 63110; USA

[32] Department of Medicine; University of California San Diego; La Jolla, CA 92093; USA

[33] Department of Computer Science and Engineering; University of California San Diego; La Jolla, CA 92093; USA

[34] Department of Bioengineering; University of California San Diego; La Jolla, CA 92093; USA

## Acknowledgments

Funding sources: We thank the CCF 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, Inc. through reagent donation and by Danone Nutricia Research in partnership with the Center for Microbiome Innovation at UC San Diego. We would like to thank Erfan Sayyari, Dominic S. Nguyen, Elaine Wolfe and Karenina Sanders for sample processing, and Jeff DeReus for data handling, processing and maintaining the computational infrastructure. JPS was supported by SD IRACDA (5K12GM068524–17), and in part by USDA-NIFA (2019–67013-29137) and the Einstein Institute GOLD project (R01MD011389). RC and MG were supported by the Krupp Endowed Fund; RC 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 RT, RHM 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 MFO and FAPESP (2014/50265–3) to NPL. DW was supported by NIH/NHLBI Training Grant (NIH T32 HL149646). KS was supported by a PROMOS fund (DAAD). WB is a postdoctoral researcher of the Research Foundation – Flanders (FWO). RJD was supported by NIH DP2 AT010401–01. We thank Ricardo 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: Daniela 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 Austin Durant for coordinating sampling at Fermentation Festivals and the wonderful staff at San Diego Zoo Wildlife Alliance for coordinating and helping with sample collection: Michele Gaffney, Edith Galindo, Katie Kerr, Andrea Fidgett, Jennifer Stuart, Debbie Tanciatco, and Lisa 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.

## References

1). Knights D, et al., Nat Methods. 2011, 8, 8761.

2). Ono H, Scientific Data, 2017, 4, 170105. [PubMed: 28850115]

3). Bono H, PloS One, 2020,15, e0227076. [PubMed: 31978081]

4). Turnbaugh PJ Nature, 2007, 449, 804 [PubMed: 17943116]

5). Haug K, et al., Nucleic Acids Research, 2020, 48, D440. [PubMed: 31691833]

6). Damen H, et al., Analytica Chimica Acta, 1978, 103 (4), 289.

7). Robin S, et al., Nature Communications, 2021, in press.

8). Li C, et al., 2021, BioRxiv doi: 10.1101/2021.01.06.425569

9). Wang M, et al., Nature Biotechnology, 2016, 34, 828.

10). Barabási A-L, et al., Nature Food, 2020, 1, 33.

11). Maruvada P, et al., Advances in Nutrition, 2020, 11, 200. [PubMed: 31386148]

12). Sprecher K, et al., Sleep, 2019, 42, zsz113. [PubMed: 31070769]

13). Scheubert K, et al., Nat Commun, 2017, 8, 1494. [PubMed: 29133785]

14). Watrous J, et al., PNAS, 2012, 109, E1743. [PubMed: 22586093]

15). Quinn R, et al., Trends Pharmacol. Sci, 2017, 38, 143. [PubMed: 27842887]

16). St. John-Williams L, et al., Scientific Data, 2019, 212, 1.

17). Sumner L, et al., Metabolomics, 2017, 3, 211.

18). Skogerson K, et al., BMC Bioinformatics, 2011, 12, 321. [PubMed: 21816034]

19). Lai Z, et al., Nature Methods, 2018, 15, 53. [PubMed: 29176591]

20). Bouslimani A, et al., PNAS, 2016, 113, E7645. [PubMed: 27849584]

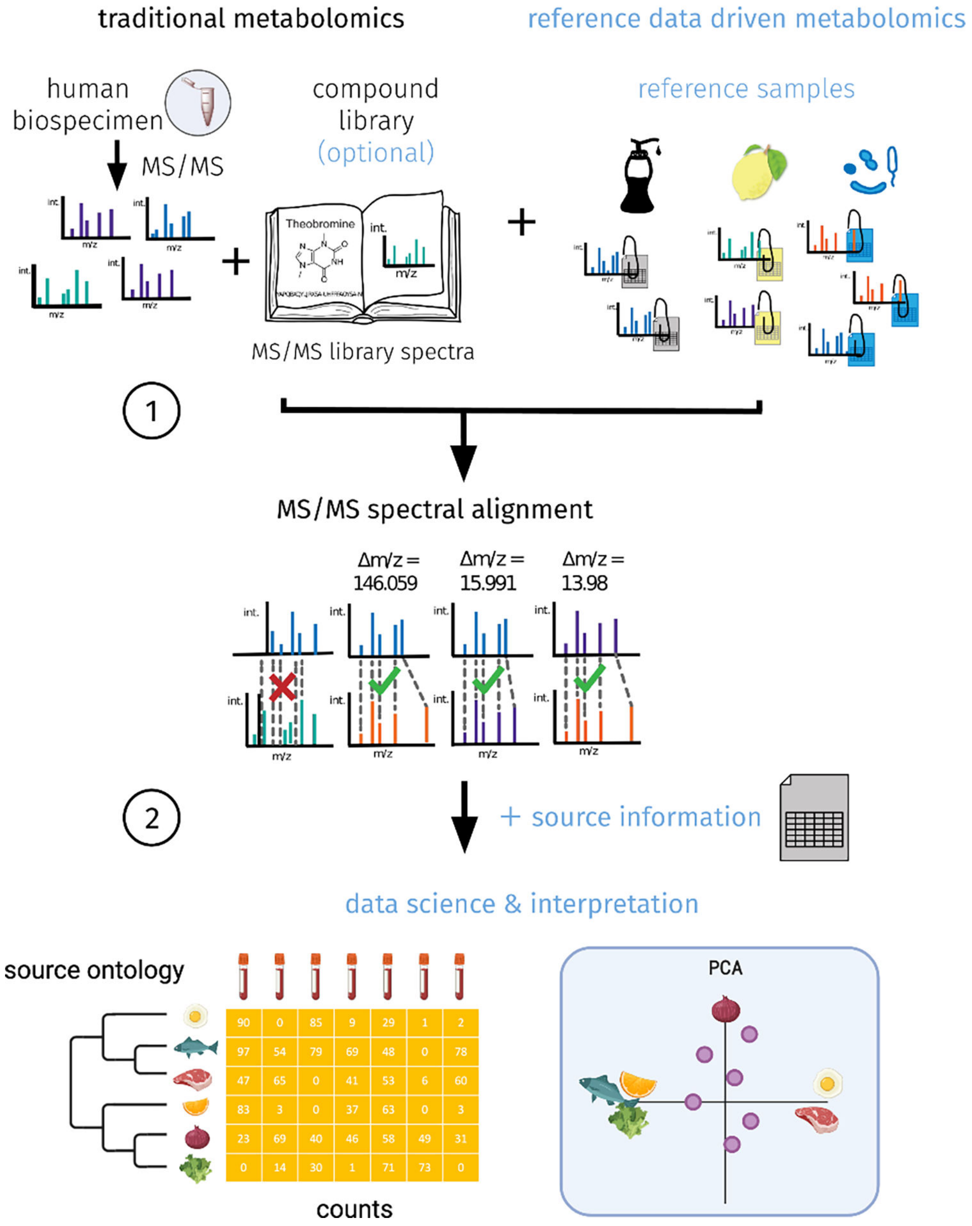
21). Wang M, et al., Nat. Biotechnology, 2020,38, 23.

22). Lungren D, et al., Expert Rev Proteomics, 2010, 7, 39. [PubMed: 20121475]

23). Tripathi T, et al., Nature Chemical Biology, 2021, 17, 146. [PubMed: 33199911]

24). Aksenov A, et al., Nature biotechnology, 2020, 39, 169.

25). West K, et al., npj Science of Food, 2022, 6, tba.



| | | | | | | |
|---|---|---|---|---|---|---|
| 90 | 0 | 85 | 9 | 29 | 1 | 2 |
| 97 | 54 | 79 | 69 | 48 | 0 | 78 |
| 47 | 65 | 0 | 41 | 53 | 6 | 60 |
| 83 | 3 | 0 | 37 | 63 | 0 | 3 |
| 23 | 69 | 40 | 46 | 58 | 49 | 31 |
| 0 | 14 | 30 | 1 | 71 | 73 | 0 |

**Figure 1. The concept of reference data-driven based analysis workflow.**

1 - 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. 2 - 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.

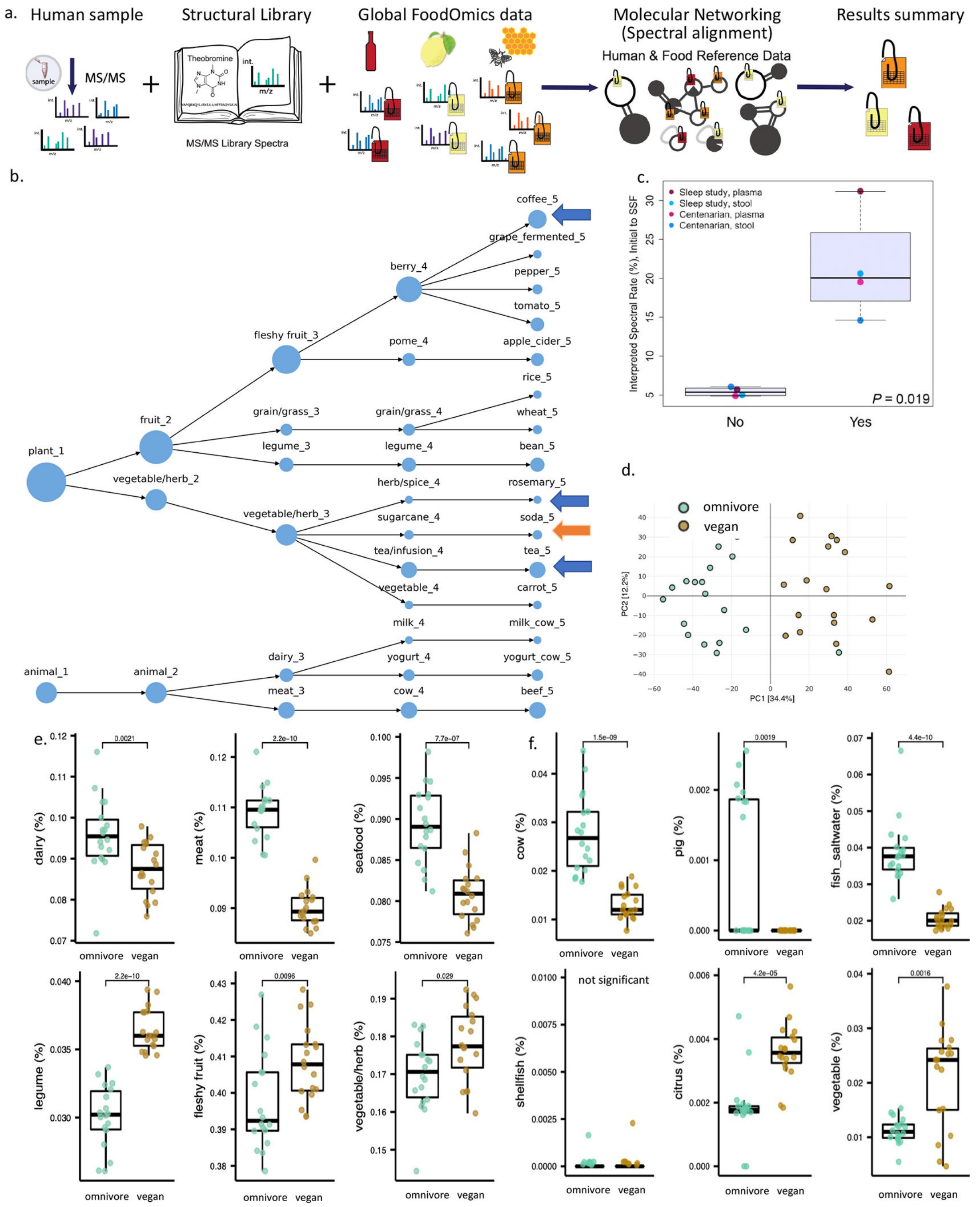


**Figure 2. RDD with food reference data.**

a. Food RDD analysis schema. b. Food spectral counts (1% FDR[13]) 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).[12] The size of node represents the relative number of spectral matches at each food level. Blue arrow - foods that could be explained based although they were not provided in the study, orange arrow– 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) vs 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, from the 25 to 75 quartile, min / max are the whiskers). d. PCA of food counts color coded by vegan (brown) vs omnivore data (green). e. Statistical analysis for the food counts at level 3 of the ontology, in relation to omnivore and vegan data (Wilcoxon test, n=36, 19 are vegan and 19 are omnivore). f. Same as e. but level 4 ontology using unique spectral counts (spectral usage is the percentage of MS/MS spectra used in the analysis. Since they are unnamed ontologies as one would find in microorganism phylogeny in microbiome science - e.g. kingdom, genus, species we have denoted these as layers, Table S1). For e-f, The boxes represent the interquartile range (IQR). Lower limit (Q1) is 25th percentile, median (Q2), upper limit (Q3) is 75th percentile. Bars show Q3+1.5xIQR and Q1–1.5xIQR.