Reference data-set driven metabolomics

Julia Gauglitz
University of California at San Diego
Kiana West
UCSD
Wout Bittremieux
University of California San Diego https://orcid.org/0000-0002-3105-1359
Candace Williams
San Diego Zoo https://orcid.org/0000-0002-1711-1183
Kelly Weldon
Center for Microbiome Innovation, University of California San Diego, La Jolla, CA
https://orcid.org/0000-0003-1064-8153
Morgan Panitchpakdi
UCSD
Francesca Diottavio
University of Teramo
Christine Aceves
UCSD
Elizabeth Brown
UCSD
Nicole Sikora
UCSD
Alan Jarmusch
UCSD
Cameron Martino
University of California, San Diego
Anupriya Tripathi
University of California San Diego
Michael Meehan
UCSD
Kathleen Dorrestein
Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego
Justin Shaffer
University of California, San Diego https://orcid.org/0000-0002-9371-6336
Roxana Coras

Division of Rheumatology, Department of Medicine, University of California San Diego
https://orcid.org/0000-0001-9547-218X

Fernando Vargas
UCSD

Lindsay DeRight Goldasich
UCSD

Tara Schwartz
UCSD

MacKenzie Bryant
University of California San Diego https://orcid.org/0000-0003-0749-2995

Greg Humphrey
Center for Microbiome Innovation and Departments of Pediatrics, Bioengineering, and Computer Science & Engineering, University of California San Diego, La Jolla, CA

Abigail Johnson
UMN

Katharina Spengler
UCSD

Pedro Belda-Ferre
Center for Microbiome Innovation and Departments of Pediatrics, Bioengineering, and Computer Science & Engineering, University of California San Diego, La Jolla, CA https://orcid.org/0000-0001-6532-1161

Edgar Diaz
UCSD

Daniel McDonald
University of California, San Diego https://orcid.org/0000-0003-0876-9060

Qiyun Zhu
UCSD

Elijah Emmanuel
UCSD

Mingxun Wang
UCSD https://orcid.org/0000-0001-7647-6097

Clarisse Marotz
Native Microbials

Kate Sprecher
UoC

Daniela Vargas RObles
Puerto Ayacucho

Dana Withrow
UoC

Gail Ackermann
UCSD
Lourdes Herrera
Wake
Barry Bradford
Michigan State University https://orcid.org/0000-0002-6775-4961
Lucas Maciel Mauriz Marques
University of São Paulo
Juliano Geraldo Amaral
Federal University of Bahia https://orcid.org/0000-0003-1823-1694
Rodrigo Moreira Silva
fmrp
Flávio Protaso Veras
fmrp
Thiago Mattar Cunha
fmrp
Rene Donizeti Ribeiro Oliveira
rmrp
Paulo Louzada-Junior
fmrp
Robert Mills
University of California, San Diego
Paulina Piotrowski
NIST
Stephanie Servetas
National Institute of Standards and Technology
Sandra DaSilva
NIST
Christina Jones
National Institute of Standards and Technology, Gaithersburg, MD
Nancy Lin
NIST
Katrice Lippa
National Institute of Standards and Technology, Gaithersburg, MD
Scott Jackson
National Institute of Standards and Technology
Rima Kaddurah Daouk
duke
Douglas Galasko

UCSD
Parambir Dulai
UCSD
Tatyana Kalashnikova
scripps
Curt Wittenberg
scripps
David Gonzalez
University of California, San Diego https://orcid.org/0000-0003-1423-5970
Robert Terkeltaub
UCSD
Megan Doty
UCSD https://orcid.org/0000-0003-2136-3719
Jae Kim
UCSD
Kyung Rhee
UCSD
Julia Beauchamp-Walters
UCSD
Kenneth Wright
UoC
Maria Gloria Dominguez-Bello
Rutgers University https://orcid.org/0000-0002-8879-6159
Mark Manery
WUSTL
Michelli Oliveira
UCSD
Brigid Boland
UCSD
Norberto Lopes
University of Sao Paulo https://orcid.org/0000-0002-8159-3658
Monica Guma
UCSD
Austin Swafford
University of California San Diego
Rachel Dutton
University of California, San Diego
Rob Knight
UCSD

Pieter Dorrestein (pdorrestein@health.ucsd.edu)
University of California San Diego https://orcid.org/0000-0002-3003-1030

Brief Communication

Keywords: metabolomics, data sets, molecular features

Posted Date: July 12th, 2021

DOI: https://doi.org/10.21203/rs.3.rs-654519/v1

Version of Record: A version of this preprint was published at Nature Biotechnology on July 7th, 2022. See the published version at https://doi.org/10.1038/s41587-022-01368-1.

Abstract

Human untargeted metabolomics studies succeed in annotating only ~10% of molecular features. We, therefore, introduce reference data-driven analysis that uses the source data as a pseudo-MS/MS reference library to match against human metabolomics MS/MS data. We demonstrate this approach with food source data, allowing an empirical assessment of dietary patterns from untargeted data but is broadly applicable and provides an additional layer of interpretability to metabolomics data.

Main

To understand the complexity of sequence data obtained in metagenomic or metatranscriptomic experiments, not only databases that contain curated genes are used to interpret the data but also reference data such as whole genomes (e.g. of microbes, viruses) or other reference sequence data sets with carefully curated metadata (e.g. developmental stage, tissue location or phenotype).[1-4] Such reference data-driven (RDD) analysis enables an increased understanding of the structure and function of complex communities by leveraging matches between genes or transcripts of known and unknown origin. By analogy, interpreting MS/MS based untargeted metabolomics data is performed by searching structural MS/MS libraries, however, leveraging reference data that includes all the known and unknown MS/MS features to further improve the insights that can be obtained from untargeted metabolomics data is not yet done.[10]

To enable RDD analysis for understanding an MS/MS-based untargeted metabolomics experiment, instead of only searching MS/MS structural libraries as has been carried out since the late 1970's [5,6], RDD now also searches against MS/MS spectra from well curated data sets (**Figure 1**). The key difference is that the output reports contextualized information obtained from source reference datasets. Source data includes MS/MS spectra of multiple ion forms of both 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 sophisticated approaches such as molecular networking. We have created a step-by-step tutorial on how to perform an RDD analysis using the GNPS ecosystem (https://ccms-ucsd.github.io/GNPSDocumentation/tutorials/rdd/).[9]

To exemplify RDD, we created a food metabolomics reference data set as there is an unmet need to retrospectively and empirically read out food and beverage information from human metabolomics data and to complement the current state-of-the-art mass spectrometry nutrition readout approaches that target up to ~150-200 metabolites.[10,11] The food reference data set consists of untargeted metabolomics and detailed metadata for ~3500 foods (**Table S1**). It contains 107,968 unique MS/MS spectra merged from a total of 1,907,765 spectra. This data accessible through GNPS and archived in the NPJ recommended repository MassIVE. Expansion of the food source data is accomplished by creation of additional data sets and deposition in GNPS/MassIVE with their metadata.

For RDD, the food source data is subjected to 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 data set[12], we were able to report if a given food category was consumed and if it agreed with the reported diet (**Figure 2b**). Of the 15 food categories, eight represented direct matches, three matches to fermented versions of the non-fermented foods consumed (e.g. yogurt instead of milk), and four categories were not documented to be consumed during the study. 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 — consistent with the elimination of caffeinated beverages in the controlled diet. This demonstrates that RDD can be used to successfully obtain diet information from untargeted metabolomics data and monitor diet adherence in controlled diet studies.

We also tested mismatched food inventories by performing a crossover with US or Italian foods to the clinical cohorts in those same regions. Crossover revealed that spectral match rates were 5–6% in the reciprocal tests, in comparison to 15–30% when the regional foods were used (**Figure 2c,** p=0.019). These observations shows that the RDD concept is applicable to metabolomics but also that RDD works optimally when the source data includes regiospecific foods.

Because RDD can be performed retrospectively, we co-analyzed the food reference dataset with 28 public human datasets (**Table S2**, **Figure 2d**). RDD increased spectral usage by 5.1±3.3 fold over structural MS/MS library matches. The inclusion of region/study specific food data significantly contributed to the increase in spectral matches (**Figure 2d;** P = 0.0028, Games-Howell test). With molecular networking, which can capture metabolized versions of molecules, spectral data usage increased by 6.8±3.5 fold. The data usage increased by 26.8±3.3% for stool data (P=2.8e-16, Games-Howell test), 27.5±5.2% for plasma data (P=0.0040) and 41±4.6% for other human data (P=0.00020). Further inclusion of connected nodes, representing potential metabolism via molecular transformations, results 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 now be leveraged as empirical readout of diet (**Figure 2d**). To assess if RDD can reveal dietary preferences, a data set of omnivores and vegans was analyzed. PCA of the spectral matches to diet revealed separation between the dietary preferences (**Figure 2e**) and that there were more MS/MS matches to dairy, meat, and seafood (P=0.0021, 2.2e-10 and 7.7e-7 respectively) in the omnivores while more MS/MS matches to legumes, fleshy fruit, and vegetables to the data from vegans (P=2.2e-10, 0.0096 and 0.029, respectively, **Figure 2f**). Because many MS/MS spectra from foods may overlap, when using only the unique MS/MS only, the results can provide additional specificity (**Figure 2g**). When performing RDD on an Alzheimer's disease population[16], it revealed that individuals with lower diet diversity consumed more dairy, sugar, soda, and coffee and that this diet type was more prevalent in the Alzheimer's dementia group. This shows that RDD can be used to retrospectively stratify clinical studies based on their diet composition.

Going forward, datasets of personal care products, medications (not just active ingredients but also formulations), microbiota, microbial isolates, etc. might also be used as source reference data. Potential applications of RDD metabolomics include understand diet and nutritional intake, medication use,

consumption of illegal substances, environmental allergens, food ingredients/adulteration, and personal care products to inform of potential exposures and health implications.

Methods, data and code availability, supporting tables 1,2 and supporting figure 3 are available as supporting information.

Declarations

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

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.

References

2) Ono, H., Scientific Data, 2017, 4, 170105.

3) Bono, H., PloS One, 2020,15, e0227076.

4) Turnbaugh P.J. Nature, 2007, 449, 804

5) Haug, K., et al., Nucleic Acids Research, 2020, 48, D440.

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: https://doi.org/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.

12) Sprecher, K., et al., Sleep, 2019, 42, zsz113.

13) Scheubert K., et al., Nat Commun., 2017, 8, 1494.

14) Watrous J., et al., PNAS, 2012, 109, E1743.

15) Quinn, R., et al., Trends Pharmacol. Sci., 2017, 38, 143.

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

Figures

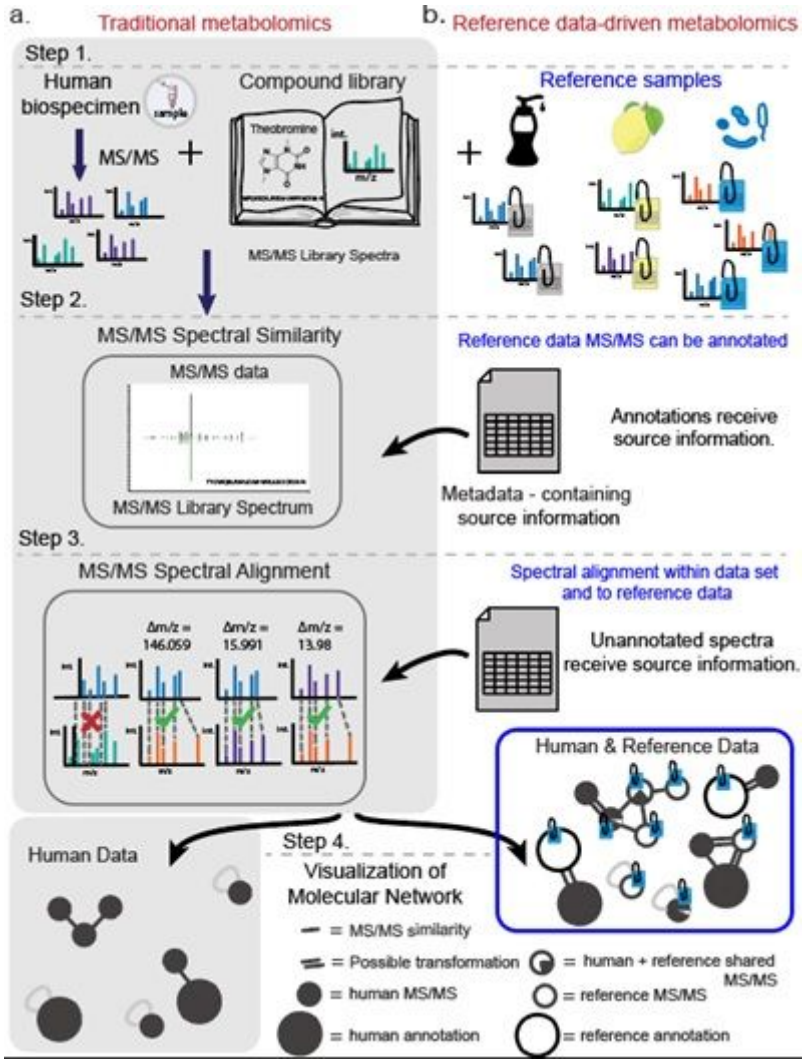

Figure 1

Retrospective reference data-driven based analysis workflow. a. Depicts the traditional untargeted mass spectrometry analysis based on structural library matching; b. Integrates the use of reference MS/MS data. RDD can leverage known and unknown mass spectrometry features. Step 1: Dataset selection; Step 2: Library search against compound-based reference libraries; Step 3: Spectral alignment to identify related or identical features - across samples; Step 4: Visualization of spectral similarity by molecular networking. Settings for spectral matching are set to ~1% FDR estimated by Passatutto.13

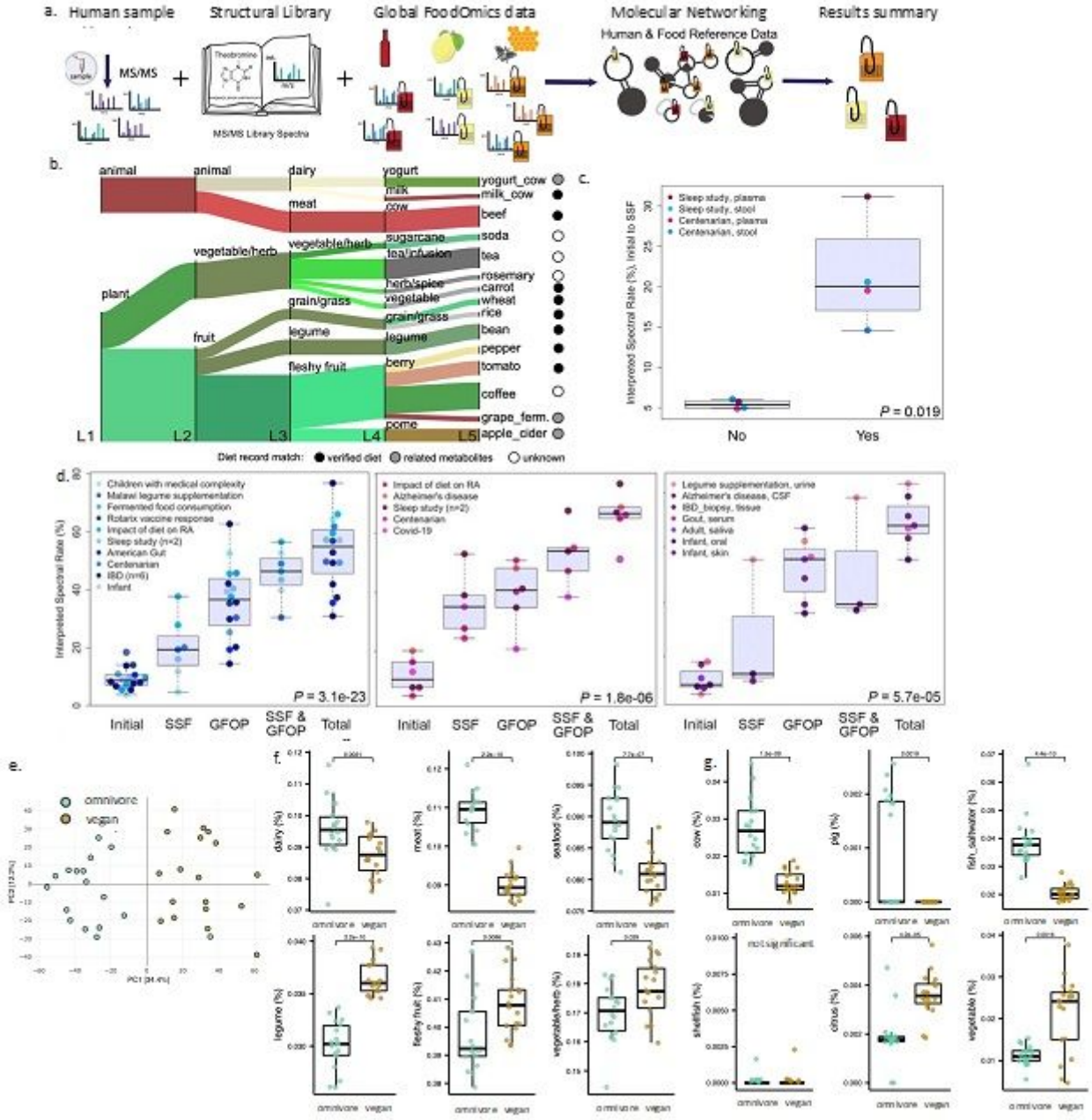

Figure 2

RDD with food reference data. a. Food RDD analysis schema. b. Food spectral counts (1% FDR13) observed in plasma from a sleep restriction and circadian misalignment study that controlled the diet of the participants.12 Solid circles represent MS/MS matches to foods consumed during the study, whereas grey circles represent MS/MS matches to fermented versions of foods consumed. 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), (Welch's t-test). d. Library spectral matches (initial), spectral matches to region or study specific foods (SSF), spectral matches to the food reference data collected via the Global FoodOmics project (GFOP), both (SSF & GFOP), expansion with molecular networking (Total). Left, stool data; middle, plasma data; right, other human biospecimens. Significant differences are determined by Welch's F-test. e. PCA of the food counts color coded by vegan (brown) vs omnivore data (green) using level 3 food ontology. f. Statistical analysis for the food spectral match level 3 ontology counts in relation to omnivore and vegan data (Wilcoxon test). g. Same as f. but level 4 ontology using unique spectral counts. The mass spectrometric analysis of animal products also contain detectable molecules from their diets resulting in spectral matches to vegan data.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- June2021SIRDDReferencedatabasedmetabolomicsNatureBiotechsubmission.docx
- TableS220200701GlobalFoodOmicsmetadataanddescriptors.xlsx
- ReportingSummary.pdf
- EditorialChecklist.pdf