BRIEF COMMUNICATION OPEN

foodMASST a mass spectrometry search tool for foods and beverages

Kiana A. West[1,2], Robin Schmid[1,2], Julia M. Gauglitz[1,2], Mingxun Wang[1,2] and Pieter C. Dorrestein[1,2,3,4]✉

There is a growing interest in unraveling the chemical complexity of our diets. To help the scientific community gain insight into the molecules present in foods and beverages that we ingest, we created foodMASST, a search tool for MS/MS spectra (of both known and unknown molecules) against a growing metabolomics food and beverage reference database. We envision foodMASST will become valuable for nutrition research and to assess the potential uniqueness of dietary biomarkers to represent specific foods or food classes.

npj Science of Food (2022)6:22 ; https://doi.org/10.1038/s41538-022-00137-3

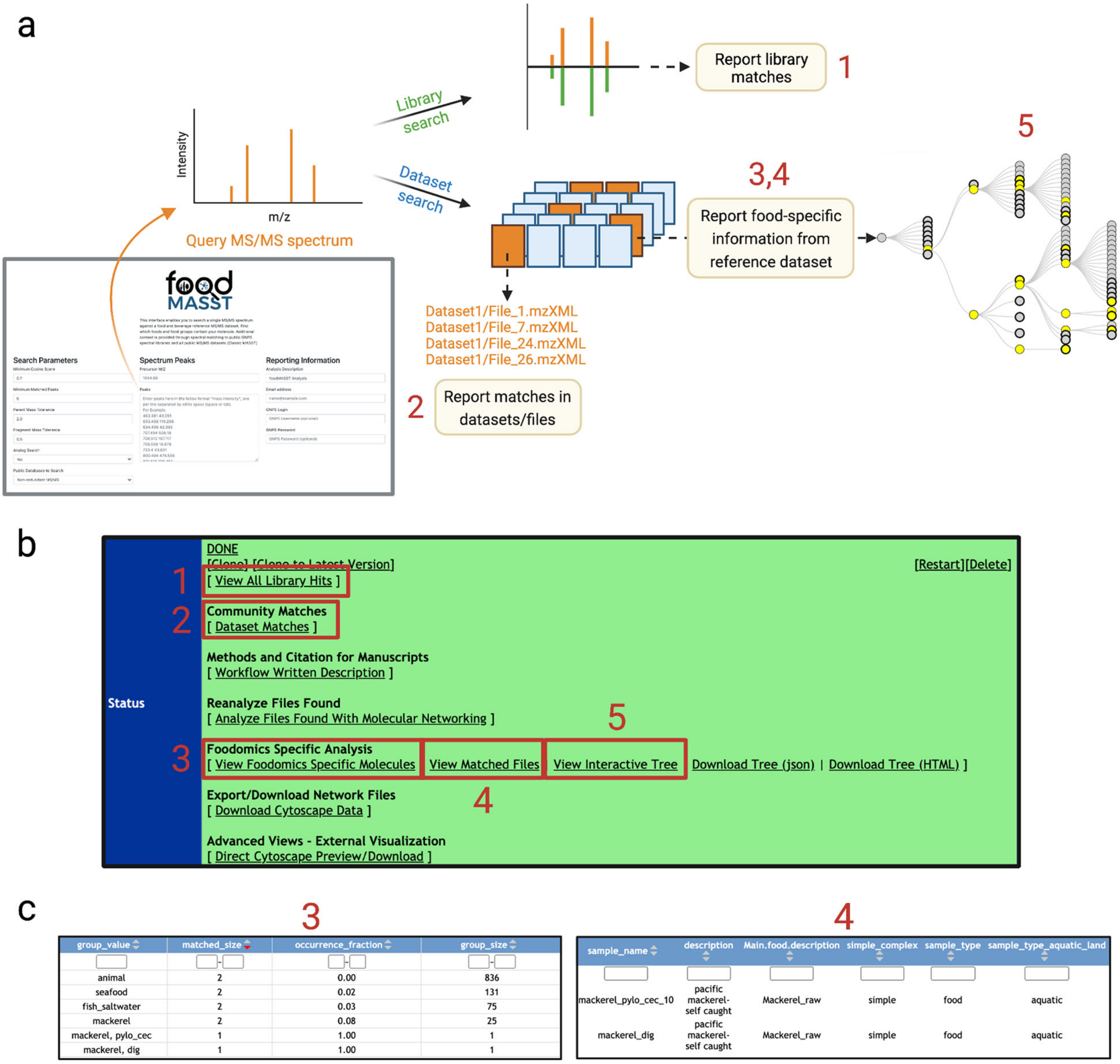
MAIN

Society is becoming more conscious of the foods we consume and more interested in understanding the compositions of specific dietary components. This sentiment is driven by the increasing awareness that diet influences our general health, including our microbiome, immune homeostasis, and even cognitive function. At the same time, our capacity to leverage untargeted metabolomics data, where many unidentified mass spectrometry signals are observed, has dramatically increased with the development of computational ecosystems such as GNPS/MassIVE[1]. There are now ~150 cataloged and tracked nutritional components, but a large proportion of the chemicals present in foods are unreported making it hard to assess if they are unique to the foods being studied[2]; it remains impossible to determine the uniqueness of these molecules across different foods and food groups.

The Mass Spectrometry Search Tool (MASST)[3] combined with a reference database of food metabolite data can be a powerful resource to understand the molecular landscape of foods. MASST is a mass spectrometry search engine that identifies all data files in the GNPS/MassIVE untargeted metabolomics repository that contain a spectral match to a query MS/MS spectrum. We created a domain-specific MASST, called foodMASST (https://masst.ucsd.edu/foodmasst), to enable reporting of the search results in the context of foods and beverages (Fig. 1a). As of Feb 2021, ~3500 untargeted metabolomics files from different foods/beverages collected as part of the Global FoodOmics Project[4] (GFOP) have been deposited in MassIVE, a public mass spectrometry repository. Each item is classified according to a customized food ontology and annotated with additional metadata (e.g., raw or cooked, cooking method, place of origin). FoodMASST utilizes this reference dataset to determine the food/beverage items containing a query spectrum and to contextualize the molecule's presence across foods. To increase usability by others in the community, we created a web interface to launch searches of known and unknown molecules with user-defined parameters and report the food information associated with the fragmentation data (MS/MS) matches. Confidence in spectral matches can be controlled by tuning the search parameters. Although powerful and useful nutrition-related search tools such as FooDB (www.foodb.ca), KNApSAcK, and others[5–8] exist, foodMASST is the only tool to enable spectral matching of both known and unknown molecules in simple (single ingredient) and complex foods and beverages.

To start a foodMASST job, the parent mass and the MS/MS spectrum details are entered into the web interface. Although the reference data that is currently searched was collected with ramped collision energies on a Q-TOF, and searching with data collected in a similar manner will give the best results, to enable searching with data that is collected differently (e.g., different collision energies, more or less background noise, and from different instruments), the data is filtered by keeping the top 6 ions in +/−50 Da window (with a 100 Dalton effective sliding window) and subjected to a square root before searching. With the cloud-based platform GNPS, all results are sharable with provenance and tied to user accounts. Once a job is completed, the results can be navigated through the landing page which displays several links to reports (Fig. 1b). To provide additional context automatic spectral library search is performed against more than 30 public spectral libraries in the GNPS/MassIVE ecosystem including GNPS contributed libraries, Human Metabolome Database[9], all three Massbanks[10–12], and many others (for a list see https://gnps.ucsd.edu/ProteoSAFe/libraries.jsp) to determine if molecules were known. "Dataset matches" navigates to all datasets (and files within those datasets) containing matches to the query spectrum. Reports specific to foodMASST can be found under "Foodomics Specific Analysis". For each category in the food ontology, the proportion of matches are reported in "View Foodomics Specific Molecules" and visualized in "View Interactive Tree". Metadata associated with the matching foods is reported in "View Matched Files". For example, when the MS/MS library spectrum for domoic acid, a potent neurotoxin from dinoflagellate blooms in the ocean[13], was searched (parent mass tolerance: 0.1 Da, ion tolerance: 0.1 Da, minimum matched peaks: 6, cosine score threshold 0.65; see data availability for job link), the only two matches to the FoodOmics reference data obtained were associated with seafood—freshly caught mackerel (Fig. 1c). It is important to consider that isomers often share similarity in their MS/MS spectra and the user should interpret the results with this in

[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]Center for Microbiome Innovation, Joan and Irwin Jacobs School of Engineering; University of California San Diego, La Jolla, CA 92093, USA. [4]Department of Pediatrics, School of Medicine, University of California San Diego, La Jolla, CA 92093, USA. ✉email: pdorrestein@ucsd.health.edu



food MASST

This interface enables you to search a single MS/MS spectrum against a food and beverage reference MS/MS dataset. Find which foods and food groups contain your molecule. Additional context is provided through spectral matching to public GNPS spectral libraries and all public MS/MS datasets (Classic MASST)

Search Parameters
Minimum Cosine Score: 0.7
Minimum Matched Peaks: 6
Parent Mass Tolerance: 2.0
Fragment Mass Tolerance: 0.5
Analog Search: No
Public Databases to Search: Non-redundant MS/MS

Spectrum Peaks
Precursor M/Z: 1044.66
Peaks: Enter peaks here in the follow format "mass intensity", one per line separated by white space (space or tab).
For Example:
463.381 43.591
693.408 119.206
694.406 42.085
707.494 506.18
708.512 197.117
709.558 18.879
723.4 43.831
800.494 476.556

Reporting Information
Analysis Description: foodMASST Analysis
Email address: name@example.com
GNPS Login: GNPS Username (optional)
GNPS Password: GNPS Password (optional)

Status

DONE
[Clone] [Clone to Latest Version]
[View All Library Hits]
[Restart][Delete]

Community Matches
[Dataset Matches]

Methods and Citation for Manuscripts
[Workflow Written Description]

Reanalyze Files Found
[Analyze Files Found With Molecular Networking]

Foodomics Specific Analysis
[View Foodomics Specific Molecules | View Matched Files | View Interactive Tree | Download Tree (json) | Download Tree (HTML)]

Export/Download Network Files
[Download Cytoscape Data]

Advanced Views - External Visualization
[Direct Cytoscape Preview/Download]

group_value	matched_size	occurrence_fraction	group_size
animal	2	0.00	836
seafood	2	0.02	131
fish_saltwater	2	0.03	75
mackerel	2	0.08	25
mackerel, pylo_cec	1	1.00	1
mackerel, dig	1	1.00	1

sample_name	description	Main.food.description	simple_complex	sample_type	sample_type_aquatic_land
mackerel_pylo_cec_10	pacific mackerel-self caught	Mackerel_raw	simple	food	aquatic
mackerel_dig	pacific mackerel-self caught	Mackerel_raw	simple	food	aquatic

Fig. 1 foodMASST workflow and reports. a A foodMASST query will search both library MS/MS spectra and metabolomics datasets that have been deposited to GNPS/MassIVE. **b** The results page lists various statistics **c** and further analysis options. Library spectra that match the query are reported (1) as potential annotations if the molecule is unknown. Public datasets, and files within datasets, containing spectral matches to the query are also reported (2) to provide context about where the molecule is observed. Additional reporting based on matches within the food reference dataset can be viewed and downloaded under "Foodomics Specific Analysis". The percentage of samples containing a spectral match for each category of the food ontology are tabulated (3). Metadata for each food item with matches is reported (4). A visual summary of the food ontology and matches in each category can be viewed in the browser (5) or downloaded.

mind. If there are multiple spectral matches possible within the scoring criteria, all spectral annotations are provided. For example, an MS/MS spectrum may match to spectra with the annotation of L-leucine, D-leucine, L-isoleucine, D-isoleucine as MS/MS-based mass spectrometry usually does not provide regiochemical and stereochemical information.

We performed additional MS/MS searches for 6 library standards and one unknown compound. These included: biocides fenamidone, spirotetramat, and enilconazole; the plant pigment cyanidin; Vitamin B_5; the antibiotic tetracycline; and an unknown molecule with a precursor *m/z* of 457.257. Fenamidone is a fungicide with low use in the US[14] and accordingly was detected in few samples (mushroom, spinach, and lettuce; Fig. 2a). For an interactive example of the results landing page see https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=16d14b8efd134fcabe227dd6377db1b9.

The "view library hits" will reveal if any of the ~620,000 public GNPS reference spectra with structural annotations match the input data. The mirror plot can be used to assess how well the query matched the reference spectrum. The "View Interactive Tree" link displays a visual representation of the food matches organized according to the GFOP ontology. Spirotetramat is an insecticide that also has low use in the US[14] and the foods sampled. However, this biocide is mainly used on citrus fruits and grapes and was detected in grapes, oranges, and cherries (Fig. 2a). Enilconazole is a fungicide mainly used on citrus fruits[14] and was detected in 31% of samples classified as citrus (Fig. 2a). Interestingly, enilconazole is also used as an antifungal in veterinary medicine and was the only of the three biocides searched that was detected in a non-plant (goat cheese) sample. A search for cyanidin (Fig. 2b), a plant pigment with reddish-purple color[15], returned teas (47% prevalence) and fruits;

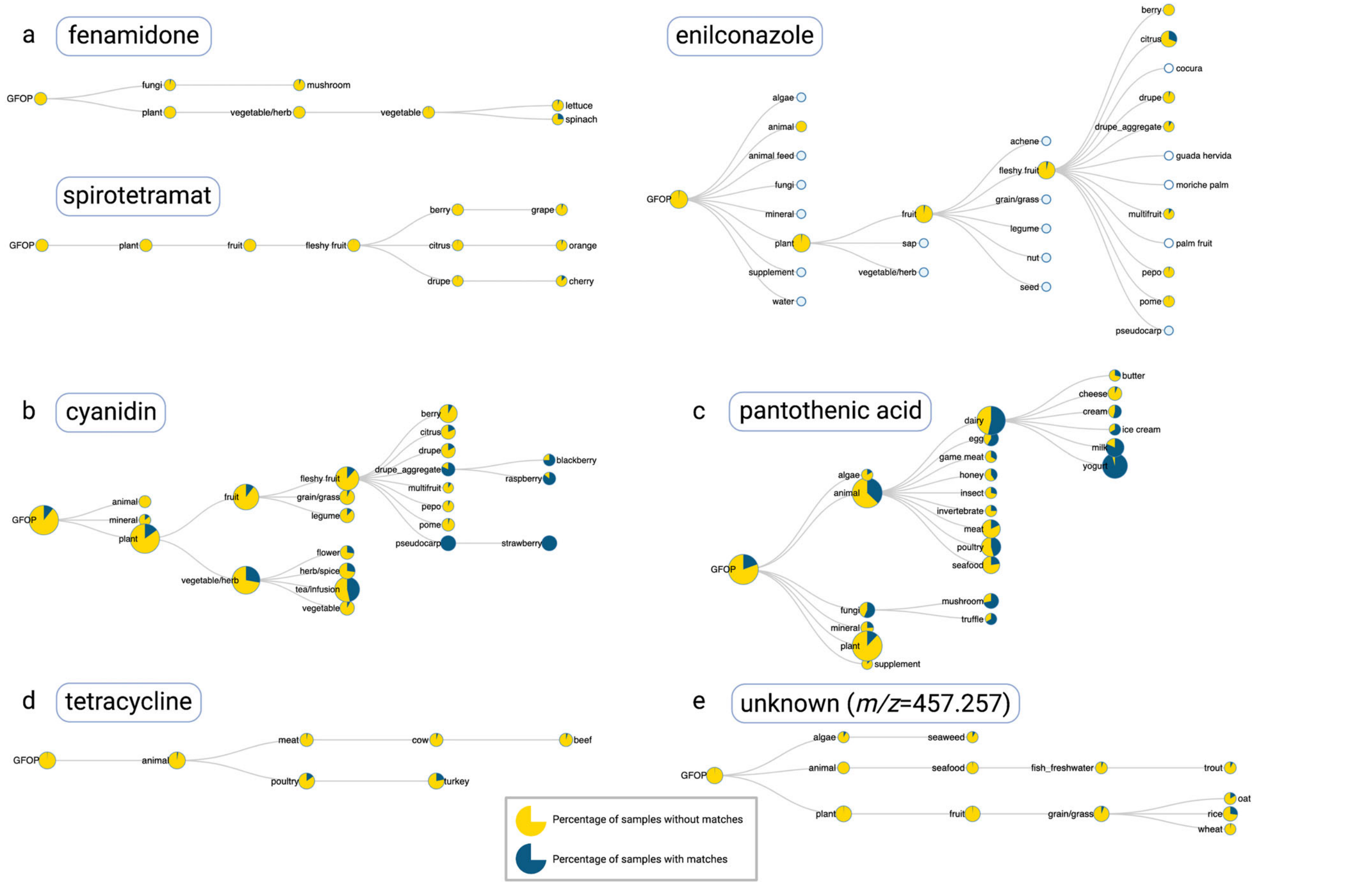


Fig. 2 Examples of foodMASST results. a Three biocides were queried against the Global FoodOmics reference dataset to determine their presence among sampled foods. **b** Cyanidin, a plant pigment responsible for reddish-purple color, was observed in the expected categories such as teas, blackberries, raspberries, and strawberries. **c** Vitamin B_5 was observed across many food categories but was most prevalent in animal and fungi samples. **d** Tetracycline, an antibiotic commonly administered to livestock, was detected only in beef and turkey samples. **e** An unknown molecule detected in biospecimens from Alzheimer's patients may be related to the consumption of oats or rice. Nodes are scaled according to the total number of samples classified for that ontology term or any of its descendants. Pies represent the percentages of samples with (blue) and without (yellow) matches to the query spectrum.

the highest prevalence was observed in raspberries (86%), blackberries (75%), and strawberries (100%). Vitamin B_5, a ubiquitous metabolite, was detected in many samples, but had the highest prevalence in animal-based foods and fungi (Fig. 2c). We also searched for an antibiotic known to be used in farmed animals. Tetracycline[16] (Fig. 2d) was detected in beef (5%) and poultry (22%). Finally, we searched for an unknown compound (Fig. 2e) that was detected in an Alzheimer's clinical cohort. The unknown had the highest detection rate in rice (27%) and oat (17%) samples, enabling the formulation that it may be associated with dietary habits. Links to the foodMASST jobs described above can be found in the data availability statement.

There are some precautions one must take to prevent the over-interpretation of the results. Additionally, there are limitations with the presented foodMASST approach that are not specific to foodMASST but rather general to MS/MS spectral matching based on untargeted metabolomics. For example, mass spectrometry can be collected in positive and negative ion modes. The reference data is currently limited to positive ion mode and thus molecules only ionizable in negative mode cannot be used, however, this infrastructure can easily accommodate negative ion mode if the community chooses to provide such reference data. Another caveat is that two different molecules, especially structurally related isomers, can have nearly identical MS/MS spectra. Another common feature of mass spectrometry is that molecules may be ionized as different adducts (e.g., H^+, Na^+, K^+, NH_4^+). It is common in an untargeted metabolomics experiment to have multiple adducts for each molecule. We encourage searching all adducts that have MS/MS information as it is impossible to get informative MS/MS matching when there are 1, 2, or 3 fragment ions. Such searches provide too little structural information to be reliable[17] and therefore the use of low information MS/MS spectral entries are discouraged. In general, the more ions and tighter the mass tolerances used for the search the less likely spurious matches are obtained.

The user may also be interested in structural analogs of related molecules in different foods as they are likely to have similar biological activities. Distributions of analogs can be discovered by searching in analog mode and reporting the neighbors of the searched spectrum in a molecular network. Analog searches will also allow improved discovery of MS/MS matches collected on different instruments or with different instrument settings.

The GFOP reference dataset will continue to grow. The community can contribute to the database that foodMASST uses by depositing LC-MS/MS-based metabolomics data, with the food-specific metadata, into GNPS/MassIVE followed by correspondence with the authors who will inspect the contributed data and add it to the existing database. We anticipate foodMASST will provide valuable insight for unknown MS/MS signals relevant to clinical studies and known signals being considered as dietary biomarkers. More broadly, the enhancement of MASST for domain-specific reporting using well-curated reference datasets will undoubtedly prove useful for many research areas.

METHODS

Reference data

The existing functionality of MASST[3] was utilized to create a workflow wherein MS/MS matches identified within the GFOP food reference dataset are reported and contextualized. The GFOP reference dataset contains untargeted metabolomics data acquired from over 3500 food and beverage samples encompassing both human and animal dietary components. Each sample is associated with metadata to describe its characteristics, source, and preparation. Samples were also organized according to a custom ontology which, at the highest level, distinguishes between plant- or animal-based foods, algae, fungi, supplements, minerals, and animal feeds. The ontology is stored and managed using WebProtégé (Stanford University, California, USA), and can be viewed at https://webprotege.stanford.edu/#projects/23f59b5a-4c29-41df-b2d8-a2ea7282d912/edit/Classes. Each sample was labeled using the most specific ontology term based on the metadata provided (e.g., one sample might be classified as a red cherry tomato while another might be broadly classified as tomato).

Reporting of spectral matches

To calculate the proportion of matches at every level of the ontology, each sample inherited all parent terms of its terminal label (e.g., terminal label: red cherry tomato; parent labels: cherry tomato, tomato, berry, fleshy fruit, fruit, plant). For every ontology term, the number of samples with that label and a spectral match to the query was divided by the total number of samples with that label. The GFOP ontology was combined with the foodMASST results and visualized in a tree structure using the D3.js library and code adapted from Rob Schmuecker (https://bl.ocks.org/robschmuecker).

Reporting summary

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

DATA AVAILABILITY

The raw data used in this study are publicly available online at MassIVE (https://massive.ucsd.edu/) under the accession MSV000084900. foodMASST jobs provided as examples can be viewed at: Domoic acid https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=d77bb006bcbf48f79d789e60c4238899; Spirotetramat https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=2784cf02f6864321a24390d00f7e3263; Enilconazole https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=f6759397ec7d4ca7995d239a1a57d2a5; Fenamidone https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=16d14b8efd134fcabe227dd6377db1b9; Cyanidin https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=b7eb71d28f52438c87d969c3e9618b9e; Pantothenic acid https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=6936ab0b89ec4a7aaec79a3a935f9e08; Tetracycline https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=e5021c2bd8cc4f41ae35da99d5bfe4f2; Unknown (m/z = 457.257) https://gnps.ucsd.edu/ProteoSAFe/status.jsp?task=da0777bdc1eb43aeb60791faf7d36011.

CODE AVAILABILITY

The open-source code for foodMASST is available on GitHub (https://github.com/mwang87/GNPS_MASST and https://github.com/CCMS-UCSD/GNPS_Workflows).

Received: 5 August 2021; Accepted: 4 March 2022;
Published online: 20 April 2022

REFERENCES

1. 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).
2. Barabási, A.-L., Menichetti, G. & Loscalzo, J. The unmapped chemical complexity of our diet. *Nat. Food* **1**, 33–37 (2019).
3. Wang, M. et al. Mass spectrometry searches using MASST. *Nat. Biotechnol.* **38**, 23–26 (2020).
4. Gauglitz, J. M. et al. Reference data based insights expand understanding of human metabolomes. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.08.194159 (2020).
5. Akimoto, N. et al. FlavonoidSearch: A system for comprehensive flavonoid annotation by mass spectrometry. *Sci. Rep.* **7**, 1–9 (2017).
6. Sakurai, N. et al. An application of a relational database system for high-throughput prediction of elemental compositions from accurate mass values. *Bioinformatics* **29**, 290–291 (2013).
7. Giacomoni, F. et al. PhytoHub, an online platform to gather expert knowledge on polyphenols and other dietary phytochemicals. (2017) - Abstract. https://hal.archives-ouvertes.fr/hal-01697081/file/2017_Giacomoni_ICPH_Qu%C3%A9bec.pdf. Accessed April 6, 2022.
8. Afendi, F. M. et al. KNApSAcK family databases: Integrated metabolite-plant species databases for multifaceted plant research. *Plant Cell Physiol.* **53**, e1 (2012).
9. Wishart, D. S. et al. HMDB 4.0: The human metabolome database for 2018. *Nucleic Acids Res.* **46**, D608–D617 (2018).
10. Horai, H. et al. MassBank: A public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
11. MoNA - MassBank of North America. https://mona.fiehnlab.ucdavis.edu/. Accessed April 6, 2022
12. MassBank Europe. https://massbank.eu/MassBank/. Accessed April 6, 2022.
13. Lefebvre, K. A. & Robertson, A. Domoic acid and human exposure risks: A review. *Toxicon* **56**, 218–230 (2010).
14. Pesticide National Synthesis Project. *U.S. Geological Survey* https://water.usgs.gov/nawqa/pnsp/usage/maps/. Accessed April 6, 2022.
15. Khoo, H. E., Azlan, A., Tang, S. T. & Lim, S. M. Anthocyanidins and anthocyanins: Colored pigments as food, pharmaceutical ingredients, and the potential health benefits. *Food Nutr. Res.* **61**, 1361779 (2017).
16. Granados-Chinchilla, F. & Rodríguez, C. Tetracyclines in food and feedingstuffs: From regulation to analytical methods, bacterial resistance, and environmental and health implications. *J. Anal. Methods Chem.* **2017**, 1315497 (2017).
17. Scheubert, K. et al. Significance estimation for large scale metabolomics annotations by spectral matching. *Nat. Commun.* **8**, 1494 (2017).

ACKNOWLEDGEMENTS

This work was made possible with the support by the CCF foundation #675191, the U.S. National Institutes of Health 1U19AG063744, R01 GM107550, P41 GM103484, and R03 CA211211. Figures created with BioRender.com.

AUTHOR CONTRIBUTIONS

P.C.D. conceptualized the idea. J.M.G. and K.A.W. organized the data and metadata. K.A.W., J.M.G., R.S., M.W., and P.C.D. performed analysis. M.W. created the website. K.A.W., R.S., and M.W. developed the code. K.A.W. and P.C.D. wrote the manuscript.

COMPETING INTERESTS

P.C.D. declares the following competing financial interests: on the scientific advisory board for Sirenas and Cybele; scientific advisor and co-founder of Enveda and Ometa LLC with approval by UC San Diego. MW declares the following competing financial interests: consultant for Sirenas and Founder of Ometa LLC. K.A.W., R.S., and J.M.G. declare no competing financial interests. All authors declare no competing non-financial interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41538-022-00137-3.

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

Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.