nature biotechnology

BRIEF COMMUNICATION

https://doi.org/10.1038/s41587-020-0700-3

Auto-deconvolution and molecular networking of gas chromatography–mass spectrometry data

Alexander A. Aksenov[1,2,42], Ivan Laponogov[3,42], Zheng Zhang[1], Sophie L. F. Doran[3], Ilaria Belluomo[3], Dennis Veselkov[4,5], Wout Bittremieux[1,2,6], Louis Felix Nothias[1,2], Mélissa Nothias-Esposito[1,2], Katherine N. Maloney[1,7], Biswapriya B. Misra[8], Alexey V. Melnik[1], Aleksandr Smirnov[9], Xiuxia Du[9], Kenneth L. Jones II[1], Kathleen Dorrestein[1,2], Morgan Panitchpakdi[1], Madeleine Ernst[1,10], Justin J. J. van der Hooft[1,11], Mabel Gonzalez[12], Chiara Carazzone[12], Adolfo Amézquita[13], Chris Callewaert[14,15], James T. Morton[15,41], Robert A. Quinn[16], Amina Bouslimani[1,2], Andrea Albarracín Orio[17], Daniel Petras[1,2], Andrea M. Smania[18,19], Sneha P. Couvillion[20], Meagan C. Burnet[20], Carrie D. Nicora[20], Erika Zink[20], Thomas O. Metz[20], Viatcheslav Artaev[21], Elizabeth Humston-Fulmer[21], Rachel Gregor[22], Michael M. Meijler[22], Itzhak Mizrahi[23], Stav Eyal[23], Brooke Anderson[24], Rachel Dutton[24], Raphaël Lugan[25], Pauline Le Boulch[25], Yann Guitton[26], Stephanie Prevost[26], Audrey Poirier[26], Gaud Dervilly[26], Bruno Le Bizec[26], Aaron Fait[27], Noga Sikron Persi[27], Chao Song[27], Kelem Gashu[27], Roxana Coras[28], Monica Guma[28], Julia Manasson[29], Jose U. Scher[29], Dinesh Kumar Barupal[30], Saleh Alseekh[31,32], Alisdair R. Fernie[31,32], Reza Mirnezami[33], Vasilis Vasiliou[34], Robin Schmid[35], Roman S. Borisov[36], Larisa N. Kulikova[37], Rob Knight[15,38,39,40], Mingxun Wang[1,2], George B. Hanna[3], Pieter C. Dorrestein[1,2,15,38] ✉ and Kirill Veselkov[3] ✉

We engineered a machine learning approach, MSHub, to enable auto-deconvolution of gas chromatography–mass spectrometry (GC–MS) data. We then designed workflows to enable the community to store, process, share, annotate, compare and perform molecular networking of GC–MS data within the Global Natural Product Social (GNPS) Molecular Networking analysis platform. MSHub/GNPS performs auto-deconvolution of compound fragmentation patterns via unsupervised non-negative matrix factorization and quantifies the reproducibility of fragmentation patterns across samples.

Given its ease of use and low operational cost, GC–MS has applications with broad societal effect, such as detection of metabolic disease in newborns, toxicology, doping, forensics, food science and clinical testing. The predominant ionization technique in GC–MS is electron ionization (EI), in which all compounds are ionized by high-energy (70-eV) electrons. Because fragmentation occurs with ionization, EI GC–MS data are subjected to spectral deconvolution, a process that separates fragmentation ion patterns for each eluting molecule into a composite mass spectrum.

The 70 eV for ionizing electrons in GC–MS has been the standard, making it possible to use decades-old EI reference spectra for annotation[1]. There are ~1.2 million reference spectra that have been accumulated and curated over a period of more than 50 years[2]. Many tools and repositories for GC–MS data have been introduced[3–15]; however, much of GC–MS data processing is restricted to vendor-specific formats and software[8]. Currently, deconvolution requires setting multiple parameters manually[3–5] or posessing computational skills to run the software[7]. Also, the lack of data sharing in a uniform format precludes data comparison between laboratories and prevents taking advantage of repository-scale information and community knowledge, resulting in infrequent reuse of GC–MS data[8,11–15].

Although batch modes exist, deconvolution quality is currently not enhanced by using information from all other files. To leverage across-file information, improve scalability of spectral deconvolution and eliminate the need for manually setting the deconvolution parameters (*m/z* error correction of the ions and peak shape—slopes of raising and trailing edges, peak RT shifts and noise/intensity thresholds), we developed an algorithmic learning strategy for auto-deconvolution (Fig. 1a–f). We deployed this functionality within GNPS/MassIVE (https://gnps.ucsd.edu)[16] (Fig. 1f–i). To promote analysis reproducibility, all GNPS jobs performed are retained in the 'My User' space and can be shared as hyperlinks.

This user-independent 'automatic' parameter optimization is accomplished via fast Fourier transform (FFT), multiplication and inverse Fourier transform for each ion across an entire data set, followed by an unsupervised non-negative matrix factorization (NMF) (one-layer neural network). Then, the compositional consistency of spectral patterns for each spectral feature deconvoluted across the entire data set can be summarized as a 'balance score'. The balance score (mathematical definition in the Methods) quantifies reproducibility of the deconvoluted fragmentation patterns across the data, which, in turn, gives insight into how well the spectral

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

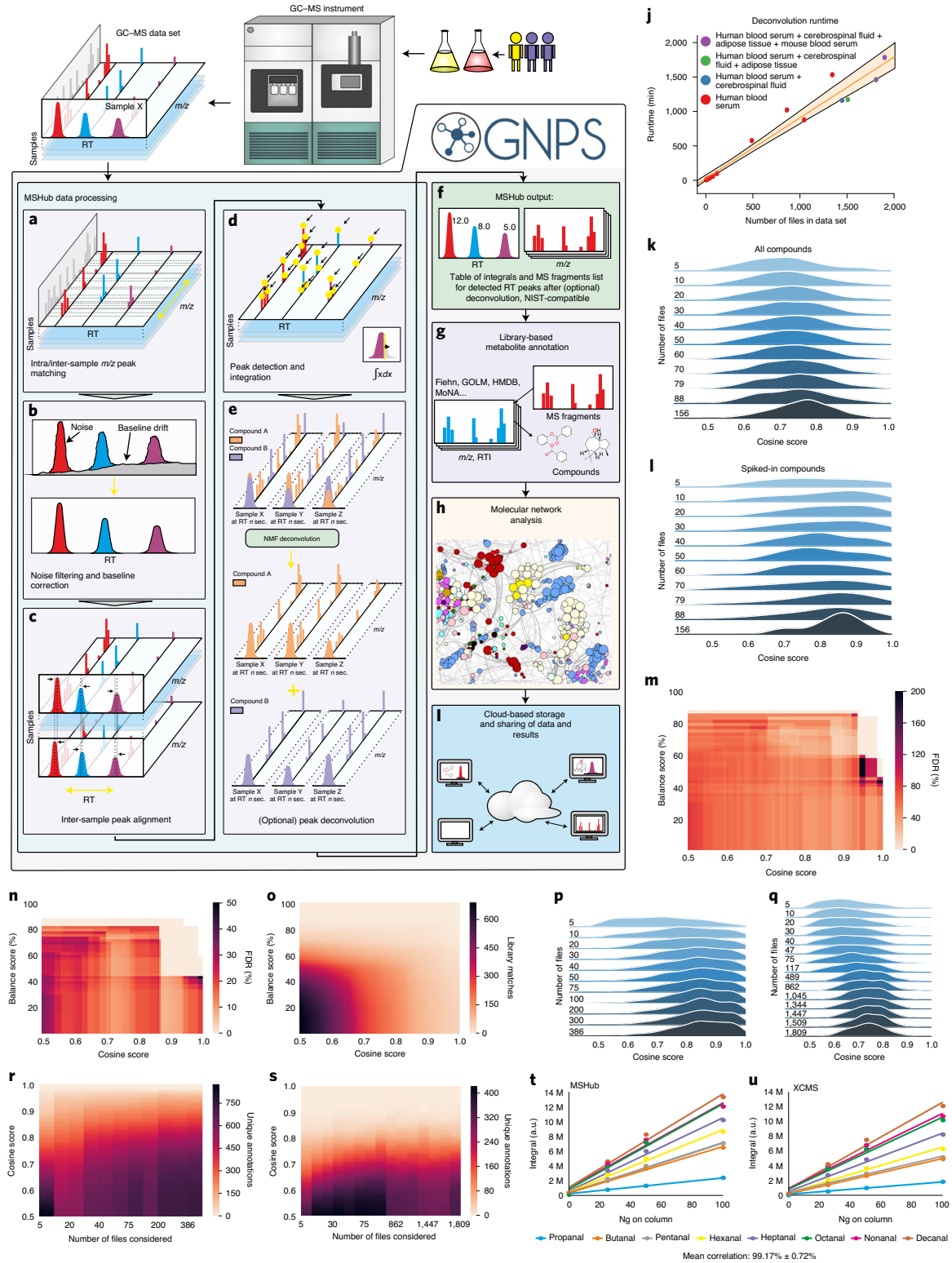

Fig. 1 | The processing pipeline and performance. a, Spectra are aligned and binned; noise is filtered and (**b**) baseline corrected. **c**, Common profile across the data set and peaks in RT dimension are aligned using FFT-accelerated correlation. **d**, Generation of both peak integrals for all samples and their common fragmentation patterns. **e**, Separation of overlapping peaks with patterns across samples using NMF. **f**, Peak integrals for all samples and canonical fragmentation patterns. NIST, National Institute of Standards and Technology. **g**, Annotation with public or private libraries. RTI, retention time index. **h**, Molecular networks. **i**, Data and results are shared between users. **j**, Linear dependence of the MSHub processing time. **k**, Distributions of library matching scores with an increased volume of data (data sets with known spiked compounds, Test1–Test11; Supplementary Table 1) for all matches and (**l**) for the spiked compounds only. **m**, FDR for annotations (Test11) of the top match and (**n**) top ten matches. **o**, Number of library matches for spiked compounds. **p**,**q**, Cosine improves as higher volume of the data enhances deconvolution quality for the top match of biological samples: breath (non-derivatized, ICL1–ICL11; Supplementary Table 1) (**p**) and human and mouse blood serum, adipose tissue and cerebrospinal fluid (silylated, data sets UCD1–UCD16; Supplementary Table 1) (**q**). **r**, The unique annotations across data sets ICL1–ICL11 and (**s**) data sets UCD1–UCD16; no balance score filtering applied. **t**,**u**, Quantitative comparison of XCMS (**t**) and MSHub (**u**). RT, retention time.

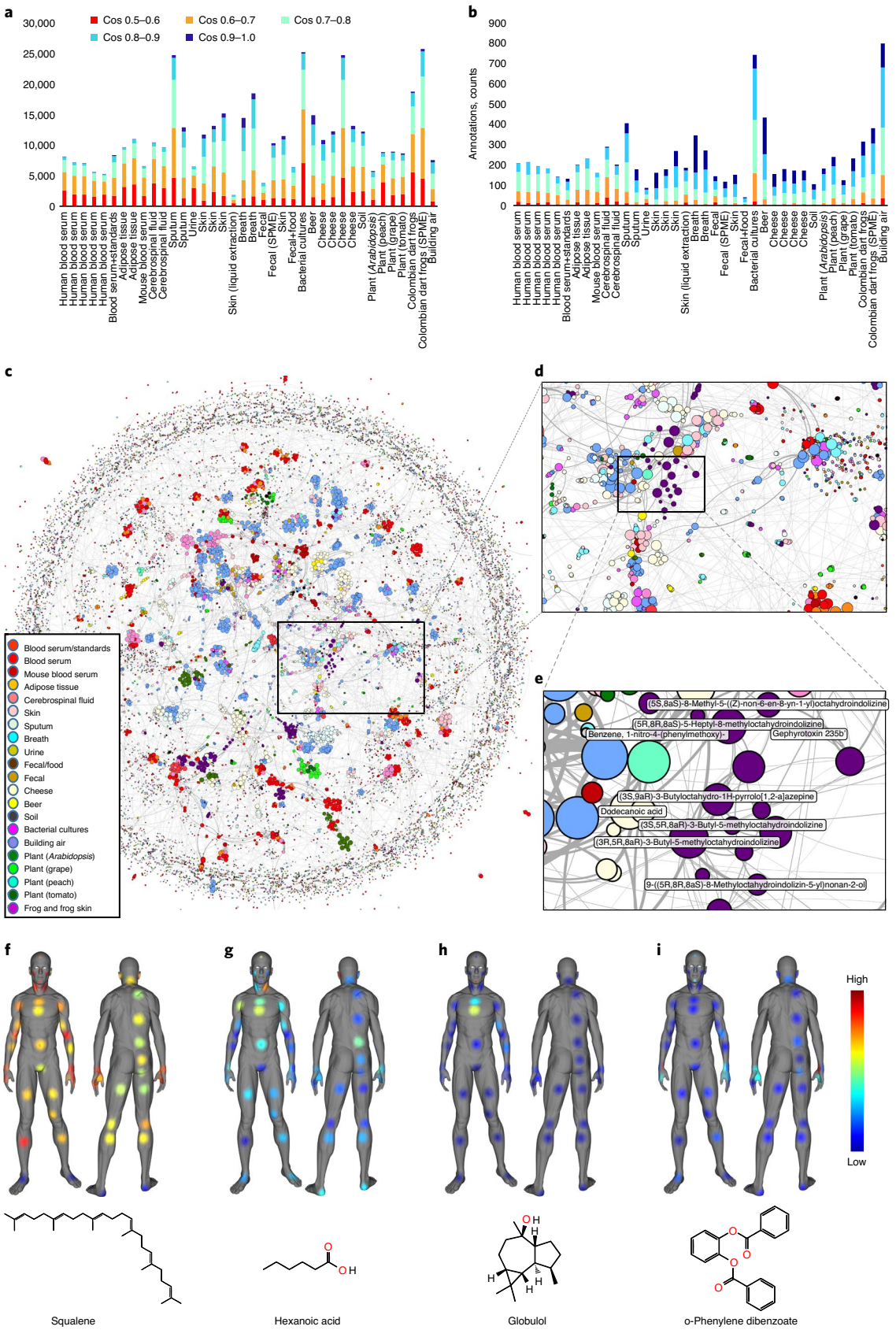

Fig. 2 | Analysis and molecular networking of GC–MS data. Annotated spectra (**a**) without filtering and (**b**) with a 65% balance score filtering. **c**, Global network containing 35,544 nodes from 8,489 files in 38 GNPS data sets. The size of the node is proportional to the number of nodes that connect, and the edge thickness is proportional to the cosine score (Supplementary Fig. 6). The annotation is the top match with cosine above 0.65. **d**, Zoomed-in region. **e**, Cluster of compounds from dart frog skin samples—all nodes are alkaloids. **f**, Human surface volatilome visualized with 'ili[19]. Molecular distributions for squalene; (**g**) hexanoic acid, a malodor molecule; (**h**) globulol, common in perfume; and (**i**) phenylene dibenzoate, common in skincare products.

feature is explained by the available data. Thus, the balance score provides an orthogonal metric of deconvoluted spectral quality. We refer to the data set spectral deconvolution tool within the GNPS environment as 'MSHub'.

All MSHub algorithms use efficient HDF5 technologies. The Fourier transform with multiplication improves MSHub's efficiency, resulting in deconvolution times that scale linearly with the number of files (Fig. 1j and Supplementary Figs. 1a, 2 and 4). We achieved this performance using out-of-core processing, a technique used to process data that are too large to fit in a computer's main memory (RAM): MSHub uploads files one at a time into the RAM module; data are then processed and deleted from memory, iteratively. Because only one sample is stored in the memory, the load is constant (Supplementary Fig. 2a–f). As machine learning approaches gain improved performance with increased volumes of information, including more data into analysis leads to better scores of spectral matches (Fig. 1k,l and Supplementary Fig. 1b). The spectral library match scores increase, and their distributions become narrower, indicating better quality of results (Fig. 1p,q). More files deconvoluted in MSHub leads to fewer chimeric spectra, resulting in higher-quality spectral features, and an increase in the number of annotations with improved scores (Fig. 1r,s). MSHub performs as well or better as other deconvolution tools (Fig. 1t,u and Supplementary Figs. 3–5). Linear scaling for MSHub makes it the only tool amenable to repository-scale operation in its present form (Supplementary Table 2). GNPS saves deconvoluted data as a summary file, so the deconvolution step does not need to be re-performed for any future analyses.

Once the summary file is generated by GNPS-MSHub or imported from another deconvolution tool, the spectra can be searched against public, private or commercial libraries. Matches are narrowed down based on user-defined filtering criteria, such as number of matched ions, Kovats index, balance score, cosine score and abundance. We provide freely available reference data of 19,808 spectra for 19,708 standards, a ~29% increase of free public libraries. All annotations should be considered level 3 (a molecular family) annotation[17]. When multiple annotations can be assigned, GNPS provides all candidate matches within the user's filtering criteria.

One of the developments that enabled finding structural relationships within mass spectrometry data is spectral alignment, which forms the basis for molecular networking[18]. GNPS has now expanded to include GC–MS-specific molecular networking[16]. GNPS-based GC–MS analysis enables data co- and re-analysis, as the processing is agnostic to the data origin. To showcase this ability, we built a global network of various public GC–MS data sets and applied a balance score of 65% (Fig. 2a,b and Supplementary Fig. 6) to ensure that only good-quality deconvoluted spectra are matched against the reference library (Fig. 2c–e and Supplementary Figs. 9 and 10). Molecular networking can further guide the annotation at the molecular family level by using information from connected nodes rather than focusing on individual annotations (Supplementary Figs. 7 and 8). One can visualize aspects such as derivatized versus non-derivatized, candidate compound class or subclass and instrument type or other metadata and inspect individual clusters of nodes (Supplementary Fig. 9). For example, we observed a cluster that belonged to dart frogs from the *Dendrobatoidea* superfamily, whereas the long-chain ketones are found in cheese and beer (Fig. 2e and Supplementary Fig. 10a). The output from GNPS can be exported for use in statistical analysis environments and for data visualization (for example, Supplementary Figs. 7–10), including molecular cartography[19] (Fig. 2f–i).

GNPS/MassIVE lowers the expertise threshold required for analysis and encourages Findable, Accessible, Interoperable and Reusable (FAIR) practices[20] by promoting re-use of GC–MS data. To highlight the broader utility of GNPS GC–MS-based analysis, videos were created (Supplemental Videos 1–6). This work aims to democratize scientific analyses. GC–MS is often the only mass spectrometry method in non-metabolomics laboratories or laboratories with fewer resources, including those in developing countries. GNPS-based GC–MS allows free access to data and reference data and to powerful computing infrastructures.

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/s41587-020-0700-3.

Received: 13 January 2020; Accepted: 9 September 2020;
Published online: 09 November 2020

References

1. Stein, S. Mass spectral reference libraries: an ever-expanding resource for chemical identification. *Anal. Chem.* **84**, 7274–7282 (2012).
2. Aksenov, A. A., da Silva, R., Knight, R., Lopes, N. P. & Dorrestein, P. C. Global chemical analysis of biology by mass spectrometry. *Nat. Rev. Chem.* **1**, 0054 (2017).
3. Smirnov, A. et al. ADAP-GC 4.0: application of clustering-assisted multivariate curve resolution to spectral deconvolution of gas chromatography–mass spectrometry metabolomics data. *Anal. Chem.* **91**, 9069–9077 (2019).
4. Tsugawa, H. et al. MS-DIAL: data-independent MS/MS deconvolution for comprehensive metabolome analysis. *Nat. Methods* **12**, 523–526 (2015).
5. Amigo, J. M., Skov, T., Bro, R., Coello, J. & Maspoch, S. Solving GC-MS problems with PARAFAC2. *Trends Anal. Chem.* **27**, 714–725 (2008).
6. Kessler, N. et al. MeltDB 2.0-advances of the metabolomics software system. *Bioinformatics* **29**, 2452–2459 (2013).
7. Domingo-Almenara, X. et al. eRah: a computational tool integrating spectral deconvolution and alignment with quantification and identification of metabolites in GC/MS-based metabolomics. *Anal. Chem.* **88**, 9821–9829 (2016).
8. Skogerson, K., Wohlgemuth, G., Barupal, D. K. & Fiehn, O. The volatile compound BinBase mass spectral database. *BMC Bioinf.* **12**, 321 (2011).
9. Akiyama, K. et al. PRIMe: a web site that assembles tools for metabolomics and transcriptomics. *In Silico Biol.* **8**, 339–345 (2008).
10. Tautenhahn, R., Patti, G. J., Rinehart, D. & Siuzdak, G. XCMS online: a web-based platform to process untargeted metabolomic data. *Anal. Chem.* **84**, 5035–5039 (2012).
11. Horai, H. et al. MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45**, 703–714 (2010).
12. Sud, M. et al. Metabolomics Workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. *Nucleic Acids Res.* **44**, D463–D470 (2016).
13. Carroll, A. J., Badger, M. R. & Harvey Millar, A. The MetabolomeExpress project: enabling web-based processing, analysis and transparent dissemination of GC/MS metabolomics datasets. *BMC Bioinf.* **11**, 376 (2010).
14. Haug, K. et al. MetaboLights—an open-access general-purpose repository for metabolomics studies and associated meta-data. *Nucleic Acids Res.* **41**, D781–D786 (2013).
15. Hummel, J. et al. Mass spectral search and analysis using the Golm Metabolome Database. in *The Handbook of Plant Metabolomics* 321–343 (Wiley, 2013).
16. 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).
17. Sumner, L. W. et al. Proposed minimum reporting standards for chemical analysis: Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* **3**, 211–221 (2007).
18. Kim, S., Gupta, N., Bandeira, N. & Pevzner, P. A. Spectral dictionaries: integrating de novo peptide sequencing with database search of tandem mass spectra. *Mol. Cell. Proteom.* **8**, 53–69 (2009).
19. Protsyuk, I. et al. 3D molecular cartography using LC–MS facilitated by Optimus and 'ili software. *Nat. Protoc.* **13**, 134–154 (2018).
20. Wilkinson, M. D. et al. The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

[1]Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California, San Diego, La Jolla, CA, USA. [2]Collaborative Mass Spectrometry Innovation Center, Skaggs School of Pharmacy and Pharmaceutical Sciences, University of California,San Diego, La Jolla, CA, USA. [3]Department of Surgery and Cancer, Imperial College London, South Kensington Campus, London, UK. [4]Intelligify Limited, London, UK. [5]Department of Computing, Imperial College, South Kensington Campus, London, UK. [6]Department of Computer Science, University of Antwerp, Antwerp, Belgium. [7]Department of Chemistry, Point Loma Nazarene University, San Diego, CA, USA. [8]Center for Precision Medicine, Department of Internal Medicine, Section of Molecular Medicine, Wake Forest School of Medicine, Winston-Salem, NC, USA. [9]Department of Bioinformatics and Genomics, University of North Carolina at Charlotte, Charlotte, NC, USA. [10]Section for Clinical Mass Spectrometry, Department of Congenital Disorders, Danish Center for Neonatal Screening, Statens Serum Institut, Copenhagen, Denmark. [11]Bioinformatics Group, Wageningen University, Wageningen, the Netherlands. [12]Department of Chemistry, Universidad de los Andes, Bogotá, Colombia. [13]Department of Biological Sciences, Universidad de los Andes, Bogotá, Colombia. [14]Center for Microbial Ecology and Technology, Ghent, Belgium. [15]Department of Pediatrics, University of California, San Diego, La Jolla, CA, USA. [16]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA. [17]IRNASUS, Universidad Católica de Córdoba, CONICET, Facultad de Ciencias Agropecuarias, Córdoba, Argentina. [18]Universidad Nacional de Córdoba, Facultad de Ciencias Químicas, Departamento de Química Biológica Ranwel Caputto, Córdoba, Argentina. [19]CONICET, Universidad Nacional de Córdoba, Centro de Investigaciones en Química Biológica de Córdoba (CIQUIBIC), Córdoba, Argentina. [20]Biological Sciences Division, Pacific Northwest National Laboratory, Richland, WA, USA. [21]LECO Corporation, St. Joseph, MI, USA. [22]Department of Chemistry and the National Institute for Biotechnology in the Negev, Ben-Gurion University of the Negev, Beer-Sheva, Israel. [23]Department of Life Sciences and the National Institute for Biotechnology in the Negev, Ben-Gurion University of the Negev, Beer-Sheva, Israel. [24]Division of Biological Sciences, University of California, San Diego, La Jolla, CA, USA. [25]UMR Qualisud, Université d'Avignon et des Pays du Vaucluse, Agrosciences, Avignon, France. [26]Laboratoire d'Etude des Résidus et Contaminants dans les Aliments (LABERCA), Oniris, INRAe, Nantes, France. [27]The French Associates Institute for Agriculture and Biotechnology of Dryland, The Jacob Blaustein Institutes for Desert Research, Ben Gurion University of the Negev, Sede Boqer Campus, Beer Sheva, Israel. [28]Division of Rheumatology, Department of Medicine, University of California, San Diego, La Jolla, CA, USA. [29]Division of Rheumatology, Department of Medicine, New York University School of Medicine, New York, NY, USA. [30]Department of Environmental Medicine and Public Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [31]Max Planck Institute for Molecular Plant Physiology, Potsdam-Golm, Germany. [32]Center of Plant Systems Biology and Biotechnology (CPSBB), Plovdiv, Bulgaria. [33]Department of Colorectal Surgery, Royal Free Hospital NHS Foundation Trust, Hampstead, London, UK. [34]Department of Environmental Health Sciences, Yale School of Public Health, Yale University, New Haven, CT, USA. [35]Institute of Inorganic and Analytical Chemistry, University of Münster, Münster, Germany. [36]A.V. Topchiev Institute of Petrochemical Synthesis RAS, Moscow, Russian Federation. [37]Peoples' Friendship University of Russia (RUDN University), Moscow, Russian Federation. [38]UCSD Center for Microbiome Innovation, University of California, San Diego, La Jolla, CA, USA. [39]Department of Bioengineering, University of California, San Diego, La Jolla, CA, USA. [40]Department of Computer Science & Engineering, University of California, San Diego, La Jolla, CA, USA. [41]Present address: Center for Computational Biology, Flatiron Institute, Simons Foundation, New York, NY, USA. [42]These authors contributed equally: Alexander A. Aksenov, Ivan Laponogov. ✉e-mail: pdorrestein@health.ucsd.edu; kirill.veselkov04@imperial.ac.uk

Methods

Tutorials and general note. The tools are accessible through http://gnps.ucsd.edu. The documentation to use the GC–MS interface can be found at https://ccms-ucsd.github.io/GNPSDocumentation/gcanalysis/.

The tutorials for the deconvolution can be accessed at https://ccms-ucsd.github.io/GNPSDocumentation/gc-ms-deconvolution/, and the library search and molecular networking instructions can be found at https://ccms-ucsd.github.io/GNPSDocumentation/gc-ms-library-molecular-network/.

The tutorial for spectral libraries upload can be found at https://ccms-ucsd.github.io/GNPSDocumentation/batchupload/.

The GNPS workflows can be launched with recommended default settings or adjusted according to user needs. The ranges and effect of settings are described in the tutorial.

The results can be inspected and quality filters applied according to user criteria.

The tutorial also describes how users can use various other aspects of GNPS functionality that include:

- Data upload and storage
- Data sharing
- Sharing analysis by sharing workflows
- Reproducing analyses
- Saving and sharing reference spectra
- Using GNPS analysis links for publishing
- Using GNPS/MassIVE repository for providing access to data along with the publication when required by the journal

The video tutorials for GNPS use for GC–MS data and examples of networking application videos can be accessed as follows:

Tutorial for the use of GNPS for analysis of GC–MS data: https://www.youtube.com/watch?v=KIOim2h8i64

GNPS for GC–MS in breathomics: using molecular networking to combine different data sets: https://www.youtube.com/watch?v=bDZj7NI-ZGw

GNPS for GC–MS in petroleomics: using molecular networks to find incorrect annotations: https://www.youtube.com/watch?v=r7DSsL03Hbk

GNPS for GC–MS in biology: using molecular networks for compound discovery in dart frogs: https://www.youtube.com/watch?v=eNLPrAjuX6w

GNPS for GC–MS in microbiology: using networks to explore the chemistry of cheese: https://www.youtube.com/watch?v=fWus3zhKbOA

GNPS for GC–MS in biochemistry: use of networking to discover antifungals produced by *Bacillus subtilis*: https://www.youtube.com/watch?v=cNPW6V3RJY4

Use of the GNPS GC–MS workflows. *GNPS GC–MS environment.* The GNPS leverages the repository infrastructure and now has expanded to include GC–MS-specific deconvolution, reference spectra matching and molecular networking tools. The new analysis workflows not only solved the scaling of analysis, but are also configured to promote data analysis reproducibility, as an analysis performed in GNPS is retained in the account-specific job tab and can be shared as a hyperlink. The user's own or someone else's shared analysis can be precisely reproduced by clicking the 'clone' button. In addition, we have enabled the community to upload and share reference spectra that then continuously accumulate, leading to continuous improvements of annotations. GNPS also gives the ability to explore all public data sets together with studies in one's private space for a particular research problem (for example, drug discovery). There are no other GC–MS deconvolution and annotation infrastructures that also work with the data in a repository. The scalability, reproducibility, capture of knowledge and ability to efficiently reuse data in the public domain make the GC–MS infrastructure in GNPS unique compared to other existing open or commercial resources. GNPS promotes FAIR use practices for mass spectrometry data[20].

The community infrastructure can be accessed at https://gnps.ucsd.edu under the header 'GC–MS EI Data Analysis'.

Deconvolution. Currently, 1D EI GC–MS data are amenable. We recommend using a minimum of ten files in the data set for deconvolution with MSHub. If the user only has fewer than ten files, spectral deconvolution and alignment should be performed using alternative methods (for example, MZmine, OpenChrom, AMDIS, MZmine/ADAP, MS-DIAL, BinBase, XCMS/XCMS Online, MetAlign, SpecAlign, SpectConnect, PARAFAC2, MeltDB or eRah). After using one of those tools, molecular networking can be performed in the same fashion as for MSHub (a detailed description is given in the Supplementary Notes), as the library search GNPS workflow accepts input from other tools into the GNPS/MassIVE environment. GNPS directly supports deconvolution output from MZmine/ADAP and MS-DIAL. The quantitative table of the deconvolution output can be used for statistical analysis with external tools.

Library search. Once the .mgf file is generated by GNPS-MSHub or imported from another deconvolution tool, the spectral features can be searched against public libraries (currently GNPS has Fiehn, HMDB, MoNA and VocBinBase) or the user's own private or commercial libraries (such as NIST and Wiley) and the freely available reference data of 19,808 spectra for 19,708 standards released with this manuscript. Users can also upload their own libraries to GNPS as well as share them with the community. Although the possible candidate annotations can be further narrowed by retention index (RI), they should still be considered level 3, a molecular family, annotation according to the 2007 Metabolomics Standards Initiative[17]. Calculation of RIs is enabled and encouraged but not enforced. When multiple annotations can be assigned, GNPS provides all candidate matches within the user's filtering criteria.

Filtering the results. The balance score is a new metric that will be available when MSHub deconvolution is used. A fragmentation pattern of a compound found to be the same in different measurements would result in a high balance score. Missing or chimeric peaks would change randomly across files and would result in a low balance score. Even when a compound is present in few samples, as long as the spectral patterns (irrespective of compound abundances) are conserved across samples, it would result in a high balance score.

Cosine and balance score should be jointly used as spectral matching filters for processing of the final results. The effect of filtering can be seen in Fig. 1m–o and Supplementary Fig. 3d,e. For the test data set shown in Fig. 1m,n, the lowest false discovery rate (FDR) of the top match is achieved with the combined threshold values of cosine >0.9 and balance score >60% (Fig. 1m). A more conservative balance score value of >80% essentially ensures the lowest observed FDR, even for poor cosine scores (here referred to as match scores). Conversely, even the high match score by itself might still result in unacceptably high FDR if the balance score is poor (Fig. 1m,n). The high match score reflects that a library spectrum exists that is similar to the query spectrum, whereas a high balance score is reflective of the high confidence in deconvolution of the spectral pattern. A well-deconvoluted pattern, as defined by the balance score, is more likely to give better matches against the spectral library. Selecting higher values of both metrics ensures that the best spectra are used and are matched to most likely annotations. The 'optimal' thresholds—that is, the values that minimize mis-annotations without being excessively restrictive—are data specific, but we recommend using the above values as a good starting point.

Molecular networks. No matter how the spectral library is searched in GC–MS, owing to the absence of a parent mass, a list of spectral matches is more likely to contain mis-annotations, both related (isomers and isobars) or, less frequently, entirely unrelated compounds[1]. However, to spot mis-assignments at the molecular family level, we propose exploring deconvoluted GC–MS data via molecular networking, a strategy that has been effective for liquid chromatography with tandem mass spectrometry (LC–MS/MS) data[16]. In the case of EI, unlike in LC–MS/MS where the precursor ion mass is known, the molecular ion is often absent. For this reason, the molecular networks are created through spectral similarity of the deconvoluted fragmentation spectrum without considering the molecular ion. We explored molecular networking patterns for the EI data (Supplementary Fig. 7) and observed that the EI-based cosine similarity networks are predominantly driven by structural similarity based on chemical class annotations (Supplementary Fig. 7a). These EI networks can be used to visualize chemical distributions and guide annotations (Supplementary Fig. 8). Some examples of molecular networking applications are discussed in the Supplemental Videos.

Three-dimensional mapping of volatilome. The sample collection and GC–MS analysis are described in the "Skin volatilome analysis" section of the Supplementary Notes. Feature tables from the deconvolution jobs for head space and liquid injection were downloaded from GNPS and combined into a single table. The coordinates for the three-dimensional model were picked for all of the sampled spots and added into the feature table as described in the tutorial (https://ccms-ucsd.github.io/GNPSDocumentation/gcanalysis/). The chemical distributors were then visualized using 'ili[19]. The chemical annotations of features have been cross-referenced from the library search jobs as described in the tutorial. Using balance filters at 50% and >0.9 cosine, we arrived at annotations that, once visualized, revealed the distributions of skin volatiles (Fig. 2f–i). For example, squalene was found on all locations but less on the feet. Hexanoic acid was most abundant on the chest and armpits. Globulol, a perfume ingredient that this individual used on the chest, was most intense on the chest, whereas phenylene dibenzoate, a skincare ingredient, was found on the face and hands.

The three-dimensional model, the feature table used for mapping and the snapshots shown in Fig. 2f–i are available at https://github.com/aaksenov1/Human-volatilome-3D-mapping-.

Generation of molecular networks. The data were collected across multiple studies as described in the Supplementary Notes. All of the data sets (Supplementary Table 1) were processed on the GNPS MSHub deconvolution workflow as described in the tutorial. The figures were generated as described in the Supplementary Notes.

Testing and validation. All modules were tested and validated individually to determine possible fail points and the results validated by manually reviewing the annotations that are obtained. The full pipeline was also tested for a variety of data sets, including those collected for this study (the 'GC–MS analysis for validation studies' section of the Supplementary Notes) and data from several previously published studies and unpublished public data. A variety of GC–MS

data are represented, including different types of mass analyzers (both high- and low-resolution instruments), different modes of sample introduction and analysis of both derivatized and non-derivatized samples. The goal was to ensure that both feature finding and library matching workflows are operational for all of these scenarios and that the results are consistent with those expected. We manually verified that the molecules that are known to be present in the data set are indeed identified and reported by the workflow. The testing information is summarized in Supplementary Table 1.

Comparison of deconvolution tools. We compared the deconvolution performance of MSHub alongside MZmine2/ADAP[3] and MS-DIAL[4]. These tools were chosen because they satisfy the following criteria: they are open, specifically designed for GC–MS data, can perform multi-file processing, are being routinely used by the metabolomics community and are actively being developed and maintained. Detailed descriptions of the procedure and parameters are given in the Supplementary Notes.

Generating input files with the alternative workflows. The MZmine2/ADAP and MS-DIAL workflows are the alternative options to perform spectral deconvolution on GC–MS data explicitly supported to be compatible with the GNPS library search workflow. For better integration, we added a new module to MZmine (version 2.52 and later) to export the quantification table (.csv) and the spectra summary file (.mgf) for the GNPS GC–MS workflow. Furthermore, a new MZmine module was also developed to enable the creation of the Kovats RI marker file compatible with the GNPS workflow. Detailed directions are given in the GNPS documentation: https://ccms-ucsd.github.io/GNPSDocumentation/gc-ms-deconvolution/.

Generation of plots. All plots were generated in Python 3.7.3, using NumPy 1.16.4, Pandas 0.25.0, RDKit 2019.03.4 and lxml 4.3.4 for data analysis purposes and Matplotlib 3.1.0 and Seaborn 0.9.0 for visualization purposes The detailed description is given in the Supplementary Notes.

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

Data availability

All of the data used in the preparation of this manuscript are publicly available at the MassIVE repository at the University of California, San Diego Center for Computational Mass Spectrometry website (https://massive.ucsd.edu). The data set accession numbers are: #1 (MSV000084033), #2 (MSV000085136), #3 (MSV000084034), #4 (MSV000084036), #5 (MSV000084032), #6 (MSV000084038), #7 (MSV000084042), #8 (MSV000084039), #9 (MSV000084040), #10 (MSV000084037), #11 (MSV000084211), #12 (MSV000083598), #13 (MSV000080892), #14 (MSV000080892), #15 (MSV000080892), #16 (MSV000084337), #17 (MSV000083658), #18 (MSV000083743), #19 (MSV000084226), #20 (MSV000083859), #21 (MSV000083294), #22 (MSV000084349), #23 (MSV000081340), #24 (MSV000084348), #25 (MSV000084378), #26 (MSV000084338), #27 (MSV000084339), #28 (MSV000081161), #29 (MSV000084350), #30 (MSV000084377), #31 (MSV000084145), #32 (MSV000084144), #33 (MSV000084146), #34 (MSV000084379), #35 (MSV000084380), #36 (MSV000084276), #37 (MSV000084277) and #38 (MSV000084212). All of the GNPS analysis jobs for all of the studies are summarized in Supplementary Table 1.

Code availability

The source code of the MSHub software, including low- and high-resolution data processing versions, is available online at Github (version used in GNPS) (https://github.com/CCMS-UCSD/GNPS_Workflows/tree/master/mshub-gc/tools/mshub-gc/proc) and at BitBucket (standalone version in MSHub developers' repository, both high and low resolution: https://bitbucket.org/iAnalytica/mshub_process/src/master/). Scripts used to parse, filter, organize data and generate the plots in the manuscript are available online at Github (https://github.com/bittremieux/GNPS_GC_fig). Script for merging individual .mgf files into a single file for creating global network is available at Github (https://github.com/bittremieux/GNPS_GC/blob/master/src/merge_mgf.py).
The three-dimensional model, the feature table with coordinates used for the mapping and the snapshots shown in Fig. 4a–d are available at https://github.com/aaksenov1/Human-volatilome-3D-mapping-. The GC–MS-adapted MolNetEnhancer code with an example Jupyter notebook can be found at https://github.com/madeleineernst/pyMolNetEnhancer. Source data are provided with this paper.

Acknowledgements

The conversion of the data from different repositories was supported by grant R03 CA211211 on reuse of metabolomics data, to build enabling chemical analysis tools for the ocean symbiosis program, and the development of a user-friendly interface for GC–MS analysis was supported by the Gordon and Betty Moore Foundation through grant GBMF7622. The University of California, San Diego Center for Microbiome Innovation supported the campus-wide seed grant awards for data collection that enabled the development of some of this infrastructure. P.C.D. was supported by the National Science Foundation (grant no. IOS-1656475) and the National Institutes of Health (NIH) (grant nos. U19 AG063744 01, P41 GM103484, R03 CA211211 and R01 GM107550). K.V. and I.L. are very grateful for the support of the Vodafone Foundation as part of the DRUGS/DreamLab project. The MSHub platform development was supported by NIH/NIAAA grant (R21 AA028432) on integrated machine learning for mass spectrometry data in liver disease, Intelligify Limited and Vodafone Foundation's DRUGS/CORONA-AI projects on network machine learning for drug repositioning and discovery of hyperfoods with antiviral/anticancer molecules. M.E. was supported by the University of Corsica. L.F.N. was supported by the NIH (R01 GM107550) and the European Union's Horizon 2020 Research and Innovation Programme (MSCA-GF, 704786). A.B. was supported by the National Institute of Justice Award (2015-DN-BX-K047). Additional support for data acquisition and data storage was provided by the Center for Computational Mass Spectrometry (P41 GM103484). The collection of data from the HomeChem Project was supported by the Sloan Foundation. G.B.H., S.L.F.D., I.L., K.V. and I.B. are grateful for the support of the OG cancer breath analysis study by the National Institute for Health Research London Invitro Diagnostic Co-operative and the NIHR Imperial Biomedical Research Centre, the Rosetrees and Stonegate Trusts and the Imperial College Charity. D.V. acknowledges support from ERC-Consolidator grant 724228 (LEMAN). I.B. acknowledges the contribution of Q. Wen and M. Colavita in the production of the training video. C. Callewaert was supported by the Research Foundation Flanders, with support from the industrial research fund of Ghent University. W.B. was supported by the Research Foundation Flanders. A.A.O. acknowledges the support of the Fulbright Commission and Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET-Argentina). The work of R.L. and P.L.B. on the data set 30 was supported by Metaboscope, part of the 'Platform 3A', funded by the European Regional Development Fund, the French Ministry of Research, Higher Education and Innovation, the Provence-Alpes-Côte d'Azur region, the Departmental Council of Vaucluse and the Urban Community of Avignon. S.A. and A.R.F. acknowledge the PlantaSYST project by the European Union's Horizon 2020 Research and Innovation Programme (SGA-CSA nos. 664621 and 739582 under FPA no. 664620). V.V. acknowledges support from the National Institute on Alcohol Abuse and Alcoholism award R24AA022057. M. Guma and R.C. acknowledge the support of the Krupp Endowed Fund grant. A portion of mass spectra in the public reference library was produced within the framework of the State Task for the Topchiev Institute of Petrochemical Synthesis RAS and with the support of the RUDN University Program 5-100. R.S.B. acknowledges support of the State Task for the Topchiev Institute of Petrochemical Synthesis RAS. L.N.K. acknowledges support of the RUDN University Program 5-100. I.M. acknowledges support of the Israel Science Foundation (project no. 1947/19) and European Research Council under the European Union's Horizon 2020 Research and Innovation Programme (project no. 640384). J.S. has been supported by NIH/NIAMS R03AR072182, the Colton Center for Autoimmunity, the Rheumatology Research Foundation, the Riley Family Foundation and the Snyder Family Foundation. J. Manasson acknowledges support from the 2017 Group for Research and Assessment of Psoriasis and Psoriatic Arthritis Pilot Research Grant and NIH/NIAMS T32AR069515. R.G. is grateful to the Azrieli Foundation for the award of an Azrieli Fellowship. J.J.J.v.d.H. acknowledges support from an ASDI eScience grant (ASDI.2017.030) from the Netherlands eScience Center-NLeSC. B.A. was supported by the National Science Foundation through the Graduate Research Fellowship Program. GC–MS analyses for collection of the MSV000083743 data set were supported by the Pacific Northwest National Laboratory, Laboratory-Directed Research and Development Program, and were contributed by the Microbiomes in Transition Initiative; data were collected in the Environmental Molecular Sciences Laboratory, a national scientific user facility sponsored by the Department of Energy (DOE) Office of Biological and Environmental Research and located at the Pacific Northwest National Laboratory (PNNL). PNNL is operated by the Battelle Memorial Institute for the DOE under contract DEAC05-76RLO1830. M. Guma and R.C. acknowledge the support of the Krupp Endowed Fund grant. R.C. was also funded by T32AR064194-07. The authors are grateful to R. da Silva for his contribution to developing the first prototype of the EI data network and his continuous assistance with further development and testing of the infrastructure. The authors are also grateful to M. Vance and D. Farmer, who organized the sampling for HomeChem Indoor Chemistry Project (https://indoorchem.org/projects/homechem/) that allowed the collection of samples for the MSV000083598 data set. B. Ross has assisted with collecting data for the MSV000084348 data set. GC–MS analyses for collection of the MSV000084211 and MSV000084212 data sets were supported by N757 Doctorados Nacionales and project EXT-2016-69-1713 from the Departamento Administrativo de Ciencia, Tecnología e Innovación (COLCIENCIAS), the seed project INV-2019-67-1747 and the FAPA project of Chiara Carazzone from the Faculty of Science at Universidad de los Andes and the grant FP80740-064-2016 of COLCIENCIAS. The authors are grateful to L. M. Garzón, P. Palacios, M. Gonzalez and J. Hernandez for their contributions to collecting the samples and to J. Oswaldo Turizo for designing and manufacturing the amphibian electrical stimulator. A.S. and X.D. acknowledge support from National Cancer Institute award U01CA235507. The authors are grateful to S. Neuman for feedback regarding the XCMS deconvolution tool.

Author contributions

P.C.D., A.A.A., M.W. and L.F.N. developed the concept of GNPS for GC–MS data. K.V. designed and supervised MSHub platform development. I.L., D.V., V.V. and

K.V. developed the MSHub platform. M.W., Z.Z. and A.A.A. developed the workflows. A.A.A., Z.Z., M.W., B.B.M. and R.S.B. performed infrastructure testing and benchmarking. A.A.A. and Z.Z. assessed EI-based molecular networking. W.B. generated plots for MSHub algorithm performance testing and benchmarking against existing deconvolution tools. Z.Z., A.A. and M.E. generated molecular network plots. M.E. and J.J.J.v.d.H. adapted the MolNetEnhancer workflow for GC–MS molecular networks. A.S., X.D., A.A.A. and B.B.M. conducted comparative testing of MSHub with existing deconvolution tools. A.A.A., A.V.M., M.P., K.L.J. and K.D. conducted three-dimensional skin volatilome mapping studies. S.L.F.D., I.B. and G.B.H. conducted the esophageal and gastric breath analysis cancers detection study. A.A.A., Z.Z., M.P. and M.W. converted and added public libraries to GNPS. A.A.A., A.V.M., S.L.F.D., C. Callewaert, B.B.M., M. Gonzalez, C. Carazzone, A.A., J.T.M., R.A.Q., A.B., A.A.O., D.P., A.M.S., S.P.C., T.O.M., M.C.B., C.D.N., E.Z., V.A., E.H.-F., R.G., M.M.M., I.M., S.E., P.L.B., B.A., R.D., R.L., Y.G., S.P., A.P., G.D., B.L.B., A.F., N.S.P., K.G., C.S., R.C., M. Guma, J. Manasson, J.U.S., D.K.B., S.A. and A.R.F. generated GC–MS data. R.S.B., L.N.K., M.P. and A.A.A. assembled the initial version of the public reference spectra library. R.S. created the MZmine export module for GNPS GC–MS input files and RI markers file export. A.A.A., R.S., I.B., A.A.O., A.M.S., B.A., M. Gonzalez, K.N.M. and R.S.B. produced training videos. M.N.-E., A.A.A., M. Gonzalez, B.B.M., A.S. and L.F.N. wrote and compiled tutorials and documentation. P.C.D., A.A.A., W.B., K.V., R.M. and R.K. wrote the paper.

Competing interests

P.C.D. is a scientific advisor for Sirenas, Galileo and Cybele. P.C.D. is scientific adviser and cofounder of Enveda and Ometa; this has been approved by UC San Diego. M.W. is a consultant for Sirenas and the founder of Ometa Labs. A.A.A. is a consultant for Ometa Labs.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41587-020-0700-3.

Correspondence and requests for materials should be addressed to P.C.D. or K.V.

Reprints and permissions information is available at www.nature.com/reprints.

nature research

Corresponding author(s): Pieter Dorrestein

Last updated by author(s): Jun 1, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Collection of GC-MS data was carried out using software supplied by the manufacturer for each individual instrument.
Data analysis	The source code of the MSHub software used for deconvolution of all data is available online at Github (version used in GNPS) (https://github.com/CCMS-UCSD/GNPS_Workflows/tree/master/mshub-gc/tools/mshub-gc/proc) and at BitBucket (standalone version in MSHub developers' repository: https://bitbucket.org/iAnalytica/mshub_process/src/master/). Scripts used to parse, filter, organize data and generate the plots in the manuscript are available online at Github (https://github.com/bittremieux/GNPS_GC_fig). Script for merging individual .mgf files into a single file for creating global network is available at Github: https://github.com/bittremieux/GNPS_GC/blob/master/src/merge_mgf.py).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All of the data used in preparation of this manuscript are publicly available at the MassIVE repository at the UCSD Center for Computational Mass Spectrometry website (https://massive.ucsd.edu). The dataset accession numbers are: #1 (MSV000084033), #2 (MSV000084033), #3 (MSV000084034), #4 (MSV000084036), #5 (MSV000084032), #6 (MSV000084038), #7 (MSV000084042), #8 (MSV000084039), #9 (MSV000084040), #10 (MSV000084037), #11 (MSV000084211), #12 (MSV000083598), #13 (MSV000080892), #14 (MSV000080892), #15 (MSV000080892), #16 (MSV000084337), #17 (MSV000083658), #18 (MSV000083743), #19 (MSV000084226), #20 (MSV000083859), #21 (MSV000083294), #22 (MSV000084349), #23 (MSV000081340), #24 (MSV000084348), #25 (MSV000084378), #26

(MSV000084338), #27 (MSV000084339), #28 (MSV000081161), #29 (MSV000084350), #30 (MSV000084377), #31 (MSV000084145), #32 (MSV000084144), #33 (MSV000084146), #34 (MSV000084379), #35 (MSV000084380), #36 (MSV000084276), #37 (MSV000084277), #38 (MSV000084212).
All of the GNPS analysis jobs for all of the studies are summarized in Table S1 provided along with the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Multiple datasets including previously collected and/or published studies were co-analyzed; the sample size in each dataset was contingent on the original study goals. The evaluation of MSHub algorithm performance was carried out for datasets of varying sizes ranging from minimum amenable of five samples to several thousands
Data exclusions	No data points were excluded from the analysis.
Replication	All analyses can be directly replicated by cloning the workflow jobs within the GNPS environment. Links to all of the jobs with source files are included in the Table S1 submitted along with the manuscript.
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	A single volunteer has been enrolled into the skin volatilome mapping study. The volunteer has had no diagnosed medical conditions, including skin disease and has been carrying out normal daily routine prior to sampling.
Recruitment	The recruitment was carried out under the UC San Diego IRB #171662.
Ethics oversight	All of the sampling for the skin volatilome mapping has been conducted under UC San Diego IRB #171662.

Note that full information on the approval of the study protocol must also be provided in the manuscript.