Supporting Information

Hydrogen rearrangement rules: computational MS/MS fragmentation and structure elucidation using MS-FINDER software

Author names

Hiroshi Tsugawa[1], Tobias Kind[2], Ryo Nakabayashi[1], Daichi Yukihira[3], Wataru Tanaka[4], Tomas Cajka[2], Kazuki Saito[1,5], Oliver Fiehn[2,6*], Masanori Arita[1,4,7*]

Author affiliations

[1]RIKEN Center for Sustainable Resource Science, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama, Kanagawa 230-0045, Japan

[2]Genome Center, University of California Davis, Davis, California 95616, USA

[3]Reifycs Inc., Minato-ku, Tokyo 105-0003, Japan

[4]Department of Genetics, SOKENDAI (The Graduate University for Advanced Studies), 1111 Yata, Mishima, Shizuoka 411-8540, Japan

[5]Graduate School of Pharmaceutical Sciences, Chiba University, 1-33 Yayoi-cho, Inage-ku, Chiba 263-8522, Japan

[6]Biochemistry Department, King Abdulaziz University, Jeddah 21589, Saudi-Arabia

[7]National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

Corresponding authors

Masanori Arita

National Institute of Genetics, 1111 Yata, Mishima, Shizuoka 411-8540, Japan

+81-55-981-9449

arita@nig.ac.jp

Oliver Fiehn

Genome Center, University of California Davis, Davis, California, USA

+1-530-754-8258

ofiehn@ucdavis.edu

Table of contents

Table S3. Internal formula database from 14 metabolome databases

Table S4. Internal structure database from 14 metabolome databases

Table S5 5,063 MassBank records used in this study

Table S6. 936 human plasma records used in this study

Table S7. Formula element statistics of formula databases to determine the practical elemental ratios

Table S8. In-house HILIC-ESI(+)-QTOFMS library of human blood plasma metabolites

Supporting Information Manuscript

MassBank and GNPS records

The MS/MS records were downloaded from MassBank (revision 173) and GNPS (downloaded on 25th March, 2016). In this study, MS/MS records were filtered by their peak *m/z* values. Only those whose *m/z* values were within 10 mDa of their theoretical mass and whose intensity exceeded 10% of their base peaks were used in MassBank records. On the other hand, the mass tolerance for GNPS was changed to 50 mDa to keep the number of available records.

The InChI or SMILES code in MassBank or GNPS records was converted to SDF files by the ChemAxon JChem molconverter (http://www.chemaxon.com). The GNPS records without a structure were not used. The exact mass was calculated with the ChemAxon JChem calculator. The theoretical precursor m/z was calculated using the exact mass and the adduct type recorded in the MassBank and GNPS field. For the records without adduct information, the adduct type was predicted by the difference in the experimental precursor m/z and the exact mass. MassBank accession CE000146, PR100831, PR100523, and EA013608 were used as the examples of reserpine, 3-indoxyl sulfate, 2'-deoxycytidine 5'-diphosphate, and guanosine 5'-diphosphoglucose, respectively. The MS/MS spectrum of *S*-1-propenylmercaptoglutathione was obtained from our previous measurements[26]. The curated MassBank records are available as **Supporting Information Table S5**

Human plasma data sets

The data of human plasma by HILIC-QTOFMS (see below) were analyzed by the MS-DIAL program[4]. Identification was performed by our in-house library (**Supporting Information Table S8**) together with manual curation with the confirmation of retention time, precursor ion, and MS/MS spectra. The mass tolerance of precursor *m/z* was set to 5 mDa. The retention time tolerance was set to 0.1 min. Other parameters of MS-DIAL were described below. The curated data of 936 records

(677 with MS/MS and 259 without MS/MS spectra) are available as **Supporting Information Table S6**.

Experimental detail of HILIC-ESI(+)-QTOFMS analysis

Six human plasma samples were obtained from the Cleveland Clinic from the GeneBank study.[30] Water, isopropanol, and acetonitrile were purchased from Fisher Optima. Ammonium formate and formic acid were purchased from Sigma–Aldrich. Authentic standards of *N*-phenylacetyl-L-glutamine and *N*6,*N*6,*N*6-trimethyl-L-lysine were obtained from Cayman Chemical Company and Sigma–Aldrich, respectively. All procedures for the metabolite extraction were kept on ice. 30 µL of human plasma was added to 1,000 µL cold mix-solvent (acetonitrile/isopropanol/water, 3:3:2, v/v/v) on ice, then vortexed for 10 s and shaken for 5 min at 4 °C using the Orbital Mixing Chilling/Heating Plate (Torrey Pines Scientific Instruments). After 2 min centrifugation at 14,000 rcf, 300 µL of the supernatant was transferred to a new 1.5 mL Eppendorf tube and evaporated to dryness in a Labconco Centrivap cold trap concentrator. The dried sample was resuspended with 60 µL (80% acetonitrile in water) and centrifuged for 5 min at 16,000 rcf. The 50 µL aliquot was transferred to a glass amber vial (National Scientific) with a micro-insert (Supelco).

The liquid chromatography system consisted of an Agilent 1290 system (Agilent Technologies Inc.) with a pump (G4220A), a column oven (G1316C) and an autosampler (G4226A). Mobile phase A was 10 mM ammonium formate with 0.125% formic acid in water; mobile phase B was 95:5 acetonitrile:water (v/v) with 10 mM ammonium formate and 0.125% formic acid. An Acquity UPLC BEH Amide column (150 × 2.1 mm; 1.7 µm) coupled to an Acquity UPLC BEH Amide VanGuard pre-column (5 × 2.1 mm; 1.7 µm) (Waters; Milford, MA, USA) was used. The gradient was 0 min, 100% B; 2 min, 100% B; 7.7 min, 70% B; 9.5 min, 40% B; 10.3 min, 30% B; 12.8 min, 100% B; 16.8 min, 100% B. The column flow rate was 0.4 mL/min, autosampler temperature was 4 °C, injection volume was 2 µL and column temperature was 45 °C.

Mass spectrometry was performed on a SCIEX TripleTOF 5600+ system (QTOF) equipped with a DuoSpray ion source. All analyses were performed at the high sensitivity mode for both TOF MS and product ion scan. The mass calibration was automatically performed using an APCI positive/negative calibration solution via a calibration delivery system (CDS). The data dependent analysis was used to obtain the MS/MS spectrum. The MS parameters were MS^1 accumulation time, 100 ms; MS^2 accumulation time, 60 ms; collision energy, 45 V; collision energy spread, 15 V; cycle time, 450 ms; mass range, *m*/*z* 55–1200; dependent product ion scan number, 5; intensity threshold, 300; exclusion time of precursor ion, 0 s; mass tolerance, 20 mDa; ignore peaks, within 6 Da; dynamic background subtraction, TRUE. The other parameters were curtain gas, 35; ion source gas 1, 50; ion source gas 2, 50; temperature, 300 °C; ion spray voltage floating, 4.5 kV; declustering potential, 100 V; RF transmission, *m*/*z* 45: 33%, *m*/*z* 130: 33% and *m*/*z* 400: 34%.

MS-DIAL parameters

The parameters of data processing for human plasma data with screenshots of MS-DIAL GUI.

As the tab delimited text

Data collection parameters

Retention time begin	0.5
Retention time end	10.5
Mass range begin	0
Mass range end	1200

Centroid parameters

MS1 tolerance	0.01
MS2 tolerance	0.1
Peak detection-based	True

Peak detection parameters

Smoothing method	LinearWeightedMovingAverage
Smoothing level	2
Minimum peak width	5
Minimum peak height	1000

Peak spotting parameters

Mass slice width	0.1

Exclusion mass list (mass & tolerance)

141.0006	0.0005
158.0262	0.0005
214.0891	0.0005
214.091	0.0005
214.0973	0.0005
215.0911	0.0005
215.0937	0.0005
216.0836	0.0005
216.0855	0.0005

Baseline correction parameters

Band width	5
Segment number	1

Deconvolution parameters

Peak consideration	Both
Sigma window value	0.001

Exclude after precursor	True

MSP file and MS/MS identification setting

MSP file	
Retention time tolerance	100
Accurate mass tolerance (MS1)	0.025
Accurate mass tolerance (MS2)	0.25
Identification score cut off	70

Text file and post identification (retention time and accurate mass based) setting

Text file	F:¥20150203_Plasma HILIC_From Tom¥MSDIAL-Plasma-HILIC-Pos-TextDB-VS2.txt
Retention time tolerance	0.1
Accurate mass tolerance	0.005
Identification score cut off	85

Adduct ion setting

[M+H]+

Alignment parameters setting

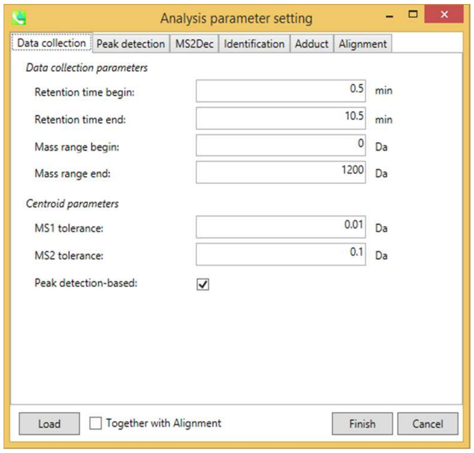
Reference file	F:¥20150203_Plasma HILIC_From Tom¥Plasma HILIC IDA¥HILIC_IDA_45CE_15CES_S01_b.abf
Retention time tolerance	0.1
MS1 tolerance	0.025
Retention time factor	0.5
MS1 factor	0.5
Peak count filter	0

QC at least filter True

Tracking of isotope labels

Tracking of isotopic labels FALSE

As the screenshots



Data collection parameters

Retention time begin: 0.5 min

Retention time end: 10.5 min

Mass range begin: 0 Da

Mass range end: 1200 Da

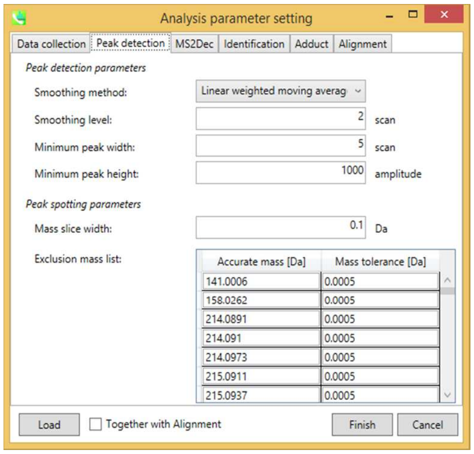
Centroid parameters

MS1 tolerance: 0.01 Da

MS2 tolerance: 0.1 Da

Peak detection-based: ☑

☐ Together with Alignment



Peak detection parameters

Smoothing method: Linear weighted moving averag

Smoothing level: 2 scan

Minimum peak width: 5 scan

Minimum peak height: 1000 amplitude

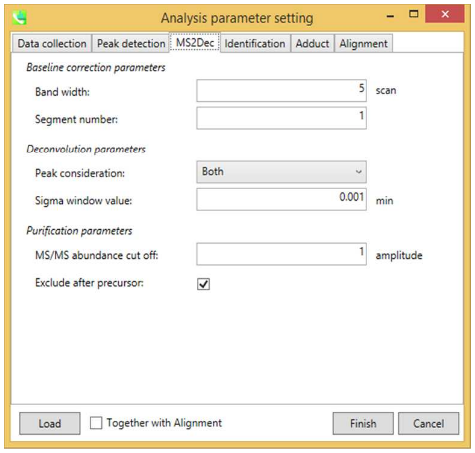
Peak spotting parameters

Mass slice width: 0.1 Da

Exclusion mass list:

Accurate mass [Da]	Mass tolerance [Da]
141.0006	0.0005
158.0262	0.0005
214.0891	0.0005
214.091	0.0005
214.0973	0.0005
215.0911	0.0005
215.0937	0.0005

☐ Together with Alignment



Baseline correction parameters

Band width: 5 scan

Segment number: 1

Deconvolution parameters

Peak consideration: Both

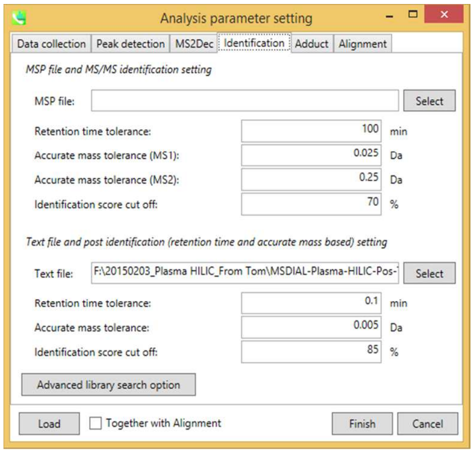
Sigma window value: 0.001 min

Purification parameters

MS/MS abundance cut off: 1 amplitude

Exclude after precursor: ☑

☐ Together with Alignment



MSP file and MS/MS identification setting

MSP file:

Retention time tolerance: 100 min

Accurate mass tolerance (MS1): 0.025 Da

Accurate mass tolerance (MS2): 0.25 Da

Identification score cut off: 70 %

Text file and post identification (retention time and accurate mass based) setting

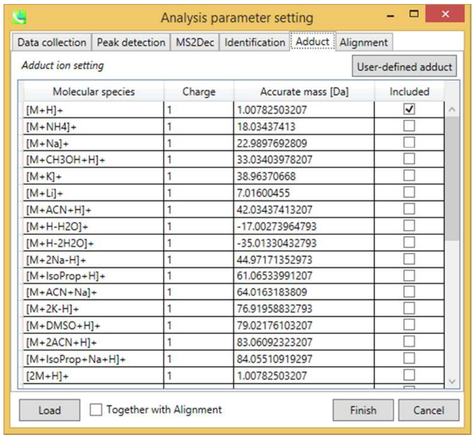
Text file: F:\20150203_Plasma HILIC_From Tom\MSDIAL-Plasma-HILIC-Pos-

Retention time tolerance: 0.1 min

Accurate mass tolerance: 0.005 Da

Identification score cut off: 85 %

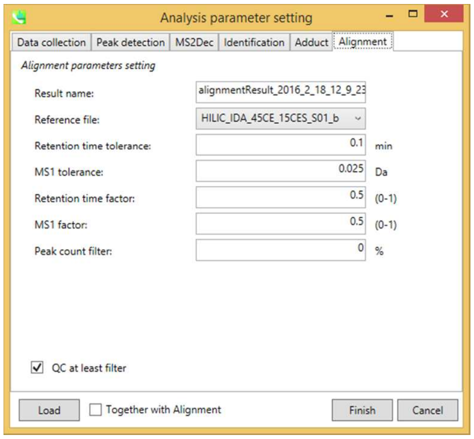
☐ Together with Alignment



Adduct ion setting

Molecular species	Charge	Accurate mass [Da]	Included
[M+H]+	1	1.00782503207	☑
[M+NH4]+	1	18.03437413	☐
[M+Na]+	1	22.9897692809	☐
[M+CH3OH+H]+	1	33.03403978207	☐
[M+K]+	1	38.96370668	☐
[M+Li]+	1	7.01600455	☐
[M+ACN+H]+	1	42.03437413207	☐
[M+H-H2O]+	1	-17.00273964793	☐
[M+H-2H2O]+	1	-35.01330432793	☐
[M+2Na-H]+	1	44.97171352973	☐
[M+IsoProp+H]+	1	61.06533991207	☐
[M+ACN+Na]+	1	64.0163183809	☐
[M+2K-H]+	1	76.91958832793	☐
[M+DMSO+H]+	1	79.02176103207	☐
[M+2ACN+H]+	1	83.06092323207	☐
[M+IsoProp+Na+H]+	1	84.05510919297	☐
[2M+H]+	1	1.00782503207	☐

☐ Together with Alignment



Alignment parameters setting

Result name: alignmentResult_2016_2_18_12_9_23

Reference file: HILIC_IDA_45CE_15CES_S01_b

Retention time tolerance: 0.1 min

MS1 tolerance: 0.025 Da

Retention time factor: 0.5 (0-1)

MS1 factor: 0.5 (0-1)

Peak count filter: 0 %

☑ QC at least filter

☐ Together with Alignment

Metabolome databases

All structure data except for STOFF were downloaded on 1st April, 2015. The STOFF data was downloaded on 1st May, 2016 from http://www.norman-network.com/?q=node/236. The SDF or CSV files of HMDB, SMPDB, YMDB, BMDB, ECMDB, FooDB, DrugBank, and T3DB were downloaded from http://www.wishartlab.com/. The SDF file of ChEBI was downloaded from https://www.ebi.ac.uk/chebi/. After the license of PlantCyc was obtained, its CSV file was downloaded via the FTP service. The MOL files of KNApSAcK were received from the developer personally. The SDF file at 'Biosystems and Pathways' in PubChem Classification Browser was also downloaded. For FooDB and PlantCyc, their SDF format was generated by ChemAxon JChem Molconverter and the MOL files of KNApSAcK were merged to create one SDF. Total 14 sets were integrated as the metabolome database in this study.

The internal formula and structure databases were prepared as follows. First, SMILES codes for all SDF records were generated by ChemAxon JChem Molconverter and were standardized by ChemAxon JChem standardizer, through stripping salt, removing fragment, removing explicit hydrogens, neutralizing, and removing absolute stereo configuration. Then, the standardized SMILES codes were converted to the 'standardized' SDF files. Exact mass, formula, and InChIKey were generated from the SDFs. The structure records with the same 14 characters of InChIKey were merged. Only organic compounds composed of C (must), H, N, O, P, S, N, F, Cl, Br, I, and Si within the range from *m/z* 50 until *m/z* 2000 were used as the internal structure database (**Supporting Information Table S4**). The internal formula database was made by integrating the structure database by their formulas (**Supporting Information Table S3**).

Structure data for program comparisons

In order to compare the function of structure elucidations with the same datasets, the structure candidates were prepared as the internal resources by the following procedure. We selected three MS/MS spectra of *N6,N6,N6*-trimethyl-L-lysine, *N*-phenylacetyl-L-glutamate, and

S-propylmercaptoglutathione as the test set. Two spectra of *N6,N6,N6*-trimethyl-L-lysine and *N*-phenylacetyl-L-glutamate were obtained by HILIC-ESI(+)-QTOFMS analysis. On the other hand, the MS/MS spectrum of *S*-propylmercaptoglutathione was obtained from our previous FT-ICRMS data[26]. The structure candidates within ±5 mDa from the experimental precursor *m/z* were retrieved from our internal databases (**Supporting Information Table S4**). Moreover, the additional candidates were downloaded from ChemSpider with the same mass tolerance in combination with the search option 'PubChem'. The reason why we used ChemSpider instead of PubChem directly was that no function for exact mass was implemented in PubChem. Finally, the numbers of structure candidates is 86, 1496, and 2368 for *N6,N6,N6*-trimethyl-L-lysine, *N*-phenylacetyl-L-glutamate, and *S*-propylmercaptoglutathione, respectively. The mass tolerances for MS1 and MS2 were set to 5 mDa (or 10 ppm) and 25 mDa (or 20 ppm) for *N6,N6,N6*-trimethyl-L-lysine and *N*-phenylacetyl-L-glutamate, respectively. The MS2 tolerance in *S*-propylmercaptoglutathione test was changed to 10 mDa (or 10 ppm).

Computer conditions and calculation of computational time

An Intel (R) Core (TM) i7-4500U CPU (1.80/2.39 GHz) with 8 GB RAM was used as Windows 10 OS. The computational times for web-based applications (CSI:FingerID, MAGMA, and MetFrag) were calculated by Google STOPWATCH. The others were calculated by C# (MS-FINDER), Python (MIDAS), and Git Bash prompt (CFM-ID).

MS-FINDER parameters

The following parameters were used in all validations: the LEWIS and SENIOR check was used, isotopic ratio tolerance was 20%, element ratio check was common range (99.7%), maximum report number for formula predictions was 100, tree depth was 2, maximum report number for structure elucidations was 100, and PubChem Online setting was set to 'only use when there is no query in the internal DBs'. In the MassBank validation, the MS/MS tolerance and the relative abundance cut off were set to 10 mDa and 2.5%, respectively. The internal data from HMDB, YMDB, PubChem (Biosystems and Pathways), SMPDB, ChEBI, PlantCyc, BMDB, KNApSAcK, FooDB, ECMDB, DrugBank, and T3DB were used for structure elucidations. In the human plasma validation, the MS/MS tolerance and the relative abundance cut off were set to 25 mDa and 2.5%, respectively. The internal data from HMDB, FooDB, and DrugBank were used. For program comparisons, the MS/MS tolerances and the relative abundance cut off was set to 25 mDa for *N6,N6,N6*-trimethyl-L-lysine and *N*-phenylacetyl-L-glutamate, 10 mDa for *S*-propylmercaptoglutathione, and 2.5%, respectively. All atomic elements and database resources were used for molecular formula predictions and structure elucidations.

CSI:FingerID parameters

The web application was performed in http://www.csi-fingerid.org/. The precursor *m/z* as described in Table 2 was added to 'Parent Mass'. The molecular formula was not set. Ionization was set to $[M+H]^+$. Chemical alphabet was set to 'CHNOPS+halogens'. Allowed mass deviation was set to 15 ppm (maximum) for *N6,N6,N6*-trimethyl-L-lysine and *N*-phenylacetyl-L-glutamate, and 10 ppm for *S*-propylmercaptoglutathione, respectively. The top formula candidate was submitted to the compound identification.

MAGMa parameters

The web application was performed in http://www.emetabolomics.org/magma. The MGF format was prepared as the queries. The structure resources were uploaded. The 'bond dissociations' and 'additional small losses' were set to 3 and 1, respectively. The 'relative (ppm)' and 'absolute (Da)' were set to 20 and 0.025 for *N6,N6,N6*-trimethyl-L-lysine and *N*-phenylacetyl-L-glutamate, and 10 and 0.01 for *S*-propylmercaptoglutathione, respectively.

MetFrag parameters

The web application was performed in http://msbi.ipb-halle.de/MetFrag/. The precursor m/z was set to ‘Parent ion’ as $[M+H]^+$. The structure resources were uploaded. The ‘Mzabs’ and ‘Mzppm’ were set to 20 and 0.025 for *N6,N6,N6*-trimethyl-L-lysine and *N*-phenylacetyl-L-glutamate, and 10 and 0.01 for *S*-propylmercaptoglutathione, respectively.

CFM-ID parameters

The installation of CFM-ID was performed according to their introduction (https://sourceforge.net/p/cfm-id/wiki/Home/). The ‘num_highest’, ‘prob_thresh’, and ‘score_type’ were set to -1, 0.001, and Jaccard, respectively. The ‘param_output0.log’ and ‘param_config.txt’ downloaded from their web site were used as the param_file and config_file, respectively. The ‘ppm_mass_tol’ and ‘abs_mass_tol’ were set to 20 and 0.025 for *N6,N6,N6*-trimethyl-L-lysine and *N*-phenylacetyl-L-glutamate, and 10 and 0.01 for *S*-propylmercaptoglutathione, respectively.

MIDAS parameters

The installation of MIDAS was performed according to their introduction (http://midas.omicsbio.org/). The ‘Default_Charge_State’ and ‘Parent_Mass_Windows’ were set to 1 and 0, respectively. The ‘Positive_Ion_Fragment_Mass_Windows’ parameter was set to ‘0, 1, 2’. The ‘Mass_Tolerance_Parent_Ion’, ‘Break_rings’, ‘Fragmentation_Depth’, and ‘Number_of_Processes’ were set to 0.005, true, 3, and 35, respectively. The ‘Mass_Tolerance_Fragment_Ions’ parameter was set to 0.025 for *N6,N6,N6*-trimethyl-L-lysine and *N*-phenylacetyl-L-glutamate, and 0.01 for *S*-propylmercaptoglutathione, respectively.

30. Wang, Z.; Klipfell, E.; Bennett, B. J.; Koeth, R.; Levison, B. S.; Dugar, B.; Feldstein, A. E.; Britt, E. B.; Fu, X.; Chung, Y.-M.; Wu, Y.; Schauer, P.; Smith, J. D.; Allayee, H.; Tang, W. H. W.; DiDonato, J. A; Lusis, A. J.; Hazen, S. L. *Nature* **2011**, *472*, 57–63.