Data Dictionary Fiehn laboratory _NIH West Coast Metabolomics Center_02-24-2019

Lipidomics by ESI QTOF MS/MS

Frequently Asked Questions:

1) Is this data normalized?

We normalize smaller data sets by using the total sum of the internal standards.For larger data sets, we use our SERRF algorithm (Systematic Error Removal by Random Forest); published in 2019 in Analytical Chemistry, DOI: 10.1021/acs.analchem.8b05592, web URL http://serrf.fiehnlab.ucdavis.edu

2) Can you give me the raw data?

We are happy to give the unnomalized data if requested to metabolomics@ucdavis.edu .

3) What are the units?

The results are given as peak intensities (peak height, counts per spectrum) for the quantification that is given in the data sheet. For lipidomics, we can perform a semi-quantification upon request using the internal standards that we measured.

4) **What is a method blank?**

Method blanks are matrix free and used as negative quality controls to evaluate contamination and background noise during sample preparation and instrument analysis

4) What software should I use to analyze my data? Which statistical tests should I do? We recommend consulting with a statistician to choose the methods best suited for your study designs. If you already have a solid understanding of statistics, you can use the assortment of methods we have compiled under http://metda.fiehnlab.ucdavis.edu . Here, diverse methods for data normalization are given (including SERRF), in addition to data transformations, univariate and multivariate data analyses, regression analyses, set enrichments statistics including ChemRICH (Barupal & Fiehn, Scientific Reports. 2017;7:14567), and network analyses MetaMapp (Barupal et al, BMC Bioinformatics. 2012;13:99). As alternative, scientists in the community use http://www.MetaboAnalyst.ca

5) How can I further interpret my statistical results?

There are very few tools to perform additional interpretations, especially for lipidomic data. We are happy to collaborate with you and help in grant proposals to establish better resources!

6) Are there any tools or software that can help me identify metabolites of interest that are not in the library? We are funded by the NIH to establish better libraries and algorithms for compound identification. We also conduct courses three times per year, teaching use of specific software https://metabolomics.ucdavis.edu/courses-and-seminars/courses , including our MS-FINDER software (Tsugawa et al Analytical Chemistry. 2016; 88:7946) in addition to software used in the community MAGMA+, MetFrag, CSI:FingerID, CFM-ID.

Glossary

UHPLC ultra high pressure liquid chromatography

ESI electrospray ionization. The method uses either negative ESI or positive ESI for negatively charged or positively charged molecules.

QTOF quadrupole time of flight mass spectrometer. The method uses data acquisition using single MS (with high-

resolution TOF); for identification purposes the mass spectrometer is operated in MS/MS mode using a quadrupole for the isolation of precursor ions, followed by collision-induced dissociation (CID) in collision cell (hexapole) with support of nitrogen, and acquiring products ions using high-resolution TOF.

MS/MS also known as tandem mass spectrometry or MS^2. After soft ionization by electrospray, the precursor (intact) charged molecules are fragmented to product ions by collision with gas atoms, usually helium, nitrogen, or argon. Fragments are then analyzed by time of flight mass spectrometry to obtain accurate mass information at high resolution.

Resolving power also called resolution. In MS, resolving power defines the ability to distinguish co-eluting masses that have the same nominal mass, but different accurate mass. For TOF instruments the mass resolving power is expressed using full width at half maximum (FWHM) definition where Δm is the peak width of a given mass peak measured (in mass units) at 50% of its height.

Mass accuracy The deviation between measured mass (accurate mass) and calculated mass (exact mass) of an ion expressed as an error value (in mDa absolute error or in ppm relative error). This parameter is important for structural interpretation allowing confirmation of the target analyte identity and the calculation of elemental composition of metabolites of unknown structure (*here*: unnamed metabolites).

MTBE methyl-tertiary butyl ether

MeOH methanol

QC quality control

AC acylcarnitines

CE cholesteryl esters

FA fatty acids

FAHFA fatty acid ester of hydroxyl fatty acid

Cer ceramides

HexCer glucosylceramides

MGDG monogalactosyldiacylglycerols

DGDG digalactosyldiacylglycerol

SQDG sulfoquinovosyl diacylglycerols

GlcADGglucuronosyldiacylglycerol

PC phosphatidyl cholines (LPC is lyso-PC, see below)

PE phosphatidyl ethanolamines (LPE is lyso-PE, see below)

PI phosphatidyl inositols

PS phosphatidyl serines

PG phosphatidyl glycerols

lyso- monoacylation of complex polar lipids at the sn1 position but not at the sn2 position

TAG triacylglycerols

DG diacylglycerols

MG monoacylglycerols

SM sphingomyelin

22:1 in lipidomic nomenclature the total number of acyl carbons (*here:* 22) and double bonds (1)

CUDA 12-[[(cyclohexylamino)carbonyl]amino]-dodecanoic acid; internal standard in the resuspension solvent (mixture of methanol : toluene, 9:1, v/v) used for quality control of the injection process. **v/v** volumetric ratio

InChI International Chemical Identifier key. Denotes the exact stereochemical and atomic description of chemicals and used as universal identifier in chemical databases.

LIPIDMAPS Identifier used in the LIPIDMAPS database.

rt retention time (minutes)

mz also m/z, or mass-to-charge ratio. In metabolomics, ions are almost exclusively detected as singly charged species.

rt_mz identifier for individual metabolites in the MassHunter Quantification method consisting of the retention time and the m/z value of specific compounds.

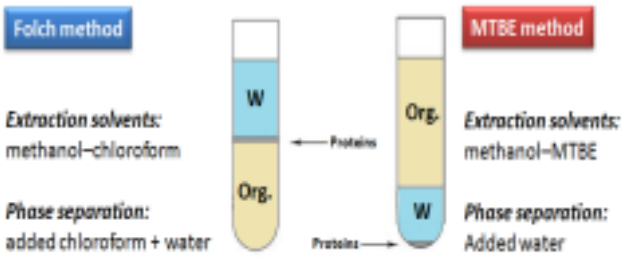
Spectral acquisition rate Time needed to obtain one mass spectrum expressed in spectra/s or Hz. In our case hundreds of primary spectra (transients) are summed and, as the final result, data acquired at 2 spectra/s are then stored in a computer.

IUPAC International Union of Pure and Applied Chemistry

NIST National Institute of Standards and Technology

istd internal standard

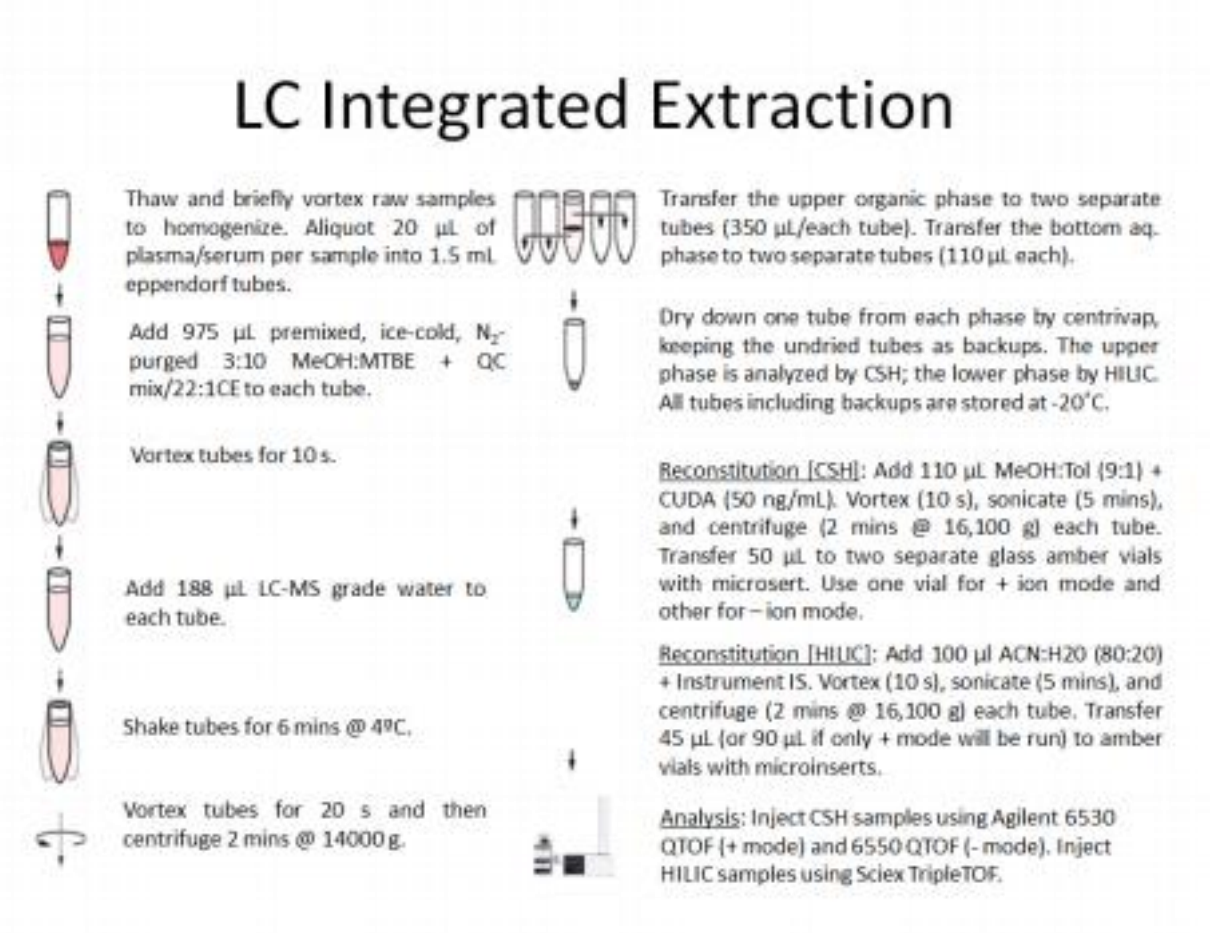
PCA Principal Component Analysis



Methods

Extraction

Blood plasma or serum is extracted following the protocols first published in Matyash V. et al., *J. Lip. Res.* **49** (2008) 1137–1146. One of the major differences to the earlier protocols by Folch or Bligh-Dyer is that in the



Matyash protocol, lipid extracts (labeled ‘org’ in the figure on the right) are separated from proteins and from polar hydrophilic small molecules (in the methanol/water phase, labeled ‘W’ in the figure below) in a way that the lipids are found in the top layer of liquid-liquid separations, rather than in the bottom layer. Decanting the top layer therefore ensures that extracts are not contaminated by proteins or polar compounds. The details of the extraction method are given in the panel to the right. We have optimized the choice of internal standards (see below) and chromatographic conditions, e.g. by using toluene in the reconstitution solvent mixture to ensure that

very lipophilic components like CE and TAGs are efficiently transferred to the UHPLC column in the injection process.

Data acquisition

Data are acquired using the following chromatographic parameters:
Column: Waters Acquity Premier BEH C18 (50 mm length x 2.1 mm internal diameter; 1.7 µm particles)
Positive Mode:
Mobile phase A: 60:40 v/v acetonitrile:water + 10 mM ammonium formate + 0.1% formic acid Mobile phase B: 90:10 v/v isopropanol:acetonitrile + 10 mM ammonium formate + 0.1% formic acid
Negative Mode:
Mobile phase A: 60:40 v/v acetonitrile:water + 10 mM ammonium acetate
Mobile phase B: 90:10 v/v isopropanol:acetonitrile + 10 mM ammonium acetate
Column temperature: 65°C
Flow-rate: 0.8 mL/min
Injection volume: 1.67 µL for ESI(+) and 5 µL for ESI(–)
Injection temperature: 4°C
Gradient: 0 min 15% (B), 0–0.75 min 30% (B), 0.75-0.98 min 48% (B), 0.98-4.00 min 82% (B), 4.00-4.13 min 99% (B), 4.13-4.50 min 99% (B), 4.50-4.58 min 15% (B), 4.58-5.50 min 15% (B)
ESI capillary voltage: ESI(+): +3.5 kV; ESI(–): –3.5 kV
Precursor/product isolation width 4 Da
Collision energy: 25 eV for ESI(+); 25 eV for ESI(–)

Scan range positive mode: m/z 120 – 1200 Da
Scan range negative mode: m/z 60–1200 Da
Spectral acquisition speed: 2 spectra/s
Mass resolution: 10,000 for ESI(+) on an Agilent 6530 QTOF MS;
20,000 for ESI(–) on an Agilent 6550 QTOF MS

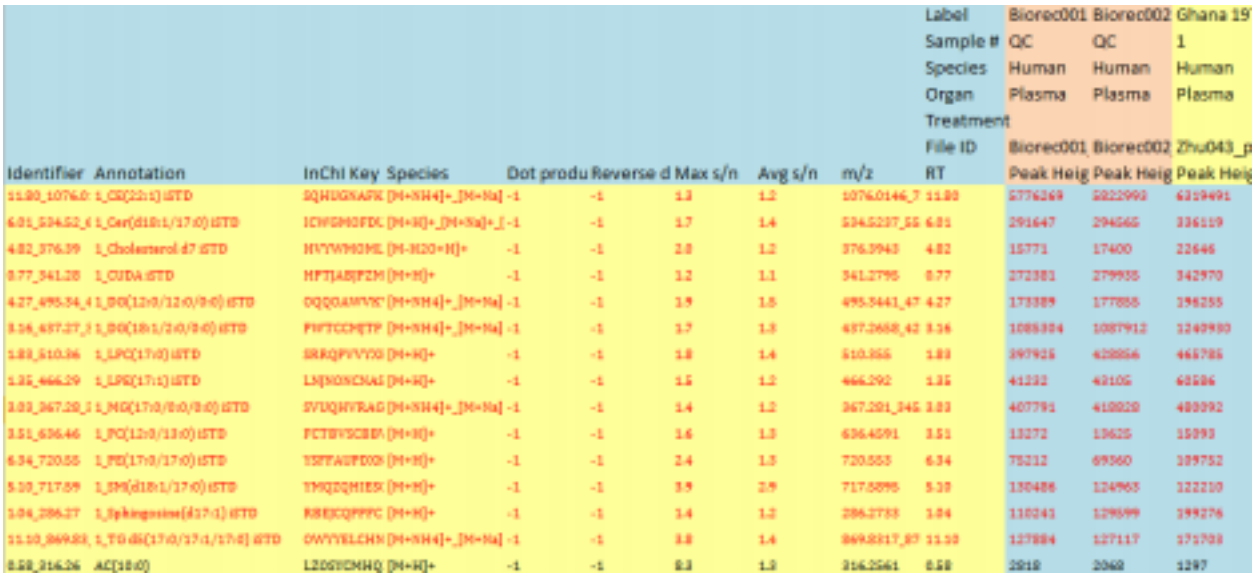
The analytical UHPLC column is protected by a short guard cartdrige (see left panel) which is replaced after 400 injections while the UHPLC column is replaced after 1,200 serum (or plasma) extract injections. We have validated that at this sequence of column replacements, no detrimental effects are detected with respect to peak shapes, absolute or relative lipid retention times or reproducibility of quantifications. This chromatography method yields excellent retention and separation of lipid classes (e.g. PC, lysoPC, PE, PS, TAG, ceramides) with narrow peak widths of 8–17 s and very good within-series retention time reproducibility of better than 6 s absolute deviation of retention times. We use automatic valve switching after each injection which we could show to reduce sample carryover for highly lipophilic compounds such as TAGs from 29% to 0.1%. This valve switching employs a dual solvent wash, first with a water/acetonitrile mixture (1:1, v/v) and subsequently with a 100% isopropanol wash.

Data processing

The general workflow for data processing is using MS-DIAL (Tsugawa et al, Nature Methods 2015;12:523), followed by a blank subtraction in Microsoft Excel and cleanup of data using MS-FLO (DeFelice et al, Analytical Chemistry 2017; 89:3250). The first step is to converted files using the Abf Converter. Default parameters are used for the processing of MS-Dial data, except for minimum peak height and width which is adjusted for the instrument where the samples ran. Once the results have been exported from MS-DIAL, a blank reduction is done based on the max peak height relative to blank average height, the average of all non-zero peak heights for samples, and if the feature is found in at least one sample. Next using MS-FLO, potential duplicates and isotopes are checked and deleted if confirmed. Then we check the MS/MS spectra before combining adducts.

Peaks are annotated in manual comparison of MS/MS spectra and accurate masses of the precursor ion to spectra

given in the Fiehn laboratory's LipidBlast spectral library (Kind et al, Nature Methods 2013; 10:755). Additional peaks were found by manual curation of sample chromatograms on a scan-by-scan basis. MassHunter Quant software was then used to verify peak candidates based on peak shape, peak height reproducibility and retention time reproducibility in replicate samples. Valid and reproducible peaks were analyzed by targeted MS/MS with the aim of increasing overall peak annotations in both positive and negative modes. These manually curated compounds are incorporated into a .txt file that has a list of accurate masses and retention time for the lipidomics platform.



									Label	Biorec001	Biorec002	Ghana 19
									Sample #	QC	QC	1
									Species	Human	Human	Human
									Organ	Plasma	Plasma	Plasma
									Treatment			
									File ID	Biorec001	Biorec002	Zhu043_p
Identifier	Annotation	InChI Key	Species	Dot produ	Reverse d	Max s/n	Avg s/n	m/z	RT	Peak Heig	Peak Heig	Peak Heig
11.80_1076.0	1_CE(22:1) iSTD	SQHUGNAFIK	[M+NH4]+,[M+Na]	-1	-1	1.3	1.2	1076.0146_7	11.80	5776269	5922993	6219491
6.01_534.52_6	1_Cer(d18:1/17:0) iSTD	ICWSMOFDL	[M+H]+,[M+Na]+,[	-1	-1	1.7	1.4	534.5237_55	6.01	291647	294565	336119
4.82_376.39	1_Cholesterol d7 iSTD	HVYWMOMLI	[M-H2O+H]+	-1	-1	2.0	1.2	376.3943	4.82	15771	17400	22646
0.77_341.28	1_CUDA iSTD	MFTJAIJFZM	[M+H]+	-1	-1	1.2	1.1	341.2795	0.77	272381	279935	342970
4.27_495.34_4	1_DG(12:0/12:0/0:0) iSTD	OQQOAWVIC	[M+NH4]+,[M+Na]	-1	-1	1.9	1.6	495.3441_47	4.27	173389	177855	196255
3.16_637.27_3	1_DG(18:1/2:0/0:0) iSTD	FWTCCMQTP	[M+NH4]+,[M+Na]	-1	-1	1.7	1.3	637.2658_42	3.16	1085304	1087912	1242930
1.83_510.36	1_LPC(17:0) iSTD	IRRQPVVYJG	[M+H]+	-1	-1	1.8	1.4	510.355	1.83	397925	428856	465785
1.35_466.29	1_LPE(17:1) iSTD	LNJNSNCNAI	[M+H]+	-1	-1	1.5	1.2	466.292	1.35	41232	42105	62586
3.03_367.20_3	1_MG(17:0/0:0/0:0) iSTD	SVUQHVRAGI	[M+NH4]+,[M+Na]	-1	-1	1.4	1.2	367.201_345	3.03	407791	410020	480092
3.51_636.46	1_PC(12:0/13:0) iSTD	FCTBVSCBBA	[M+H]+	-1	-1	1.6	1.3	636.4591	3.51	13272	13625	15393
6.34_720.55	1_PE(17:0/17:0) iSTD	YSFFAUFDOB	[M+H]+	-1	-1	2.4	1.3	720.553	6.34	75212	69360	109752
5.10_717.59	1_SM(d18:1/17:0) iSTD	YMQZQMIEBX	[M+H]+	-1	-1	3.9	2.9	717.5895	5.10	130486	124965	122210
1.04_286.27	1_Sphingosine(d17:1) iSTD	RBEJCQFFFC	[M+H]+	-1	-1	1.4	1.2	286.2733	1.04	118241	129699	199276
11.10_869.83_	1_TG dS(17:0/17:1/17:0) iSTD	OWYYELCHN	[M+NH4]+,[M+Na]	-1	-1	3.8	1.4	869.8317_87	11.10	127884	127117	171703
0.58_316.26	AC(18:0)	LZDSYCMHQ	[M+H]+	-1	-1	8.3	1.3	316.2561	0.58	2818	2062	1297

Data reporting

Data are reported including metadata, see next page as example.

The **'identifier column'** denotes the unique identifier for the technology platform, given as rt_mz. It is given for both identified and unidentified metabolites in the same manner.
The **'annotation'** denotes the name of the metabolite, if the peak has been annotated. A chemical name is not a unique identifier. We use names recognized by biologists instead of IUPAC nomenclature. If a compound is annotated, it has a name, and external database identifiers such as InChI key and LIPIDMAPS ID. Annotations are based on MS/MS matching and an in-house mz/RT library.
The **'InChI key'** hashed identifier gives the unique chemical identifier defined by the IUPAC and NIST consortia. The **'Species'** designates which adduct(s) are present for the annotated feature.
The **'max s/n'** is the ratio of the maximum value in the samples compared to the average of the blanks The **'average s/n'** is the ratio of the average value in the samples compared to the average of the blanks The **'m/z'** column details the average of the m/z value for the feature in all samples.
The **'RT'** column details the average of the RT for the feature in all samples
The row '**Label'** is the 'sample label' information that was provided by the client in the WCMC Submission Form
The row '**Sample #'** is the 'suffix' information that was provided by the client in the WCMC Submission Form. The row '**Species'** is the species of specimen information provided in the WCMC Submission Form. The row '**Organ'** is the 'Specimen type' information provided in the WCMC Submission Form. The **'treatment'** row gives the 'Treatment Group' information as listed in the WCMC Submission Form. The **'File ID'** row denotes the name of the raw data file.

Data file names are dictated by the laboratory's information and management system when the sequence starts running. QTOF raw file names from the Agilent instrumentation end with .d, QE HF raw file names from the Thermo instrument end with .raw, and TripleToF raw file names from the Sciex instrument with .wiff.

In case a sample will need to be reinjected, the file name will change from e.g.

BioRec001_389688_negCSH_preSnow001.d to BioRec001_389688_negCSH_preSnow001_2.d for the second injection, BioRec001_389688_negCSH_preSnow001_3.d for the third injection and subsequent injections.

The **actual data** are given as peak heights for the quantification ion (mz value) at the specific retention time (rt value). We give peak heights instead of peak areas because peak heights are more precise for low abundant metabolites than peak areas, due to the larger influence of baseline determinations on areas compared to peak heights. Also, overlapping (co-eluting) ions or peaks are harder to deconvolute in terms of precise determinations of peak areas than peak heights. Such data files are then called 'raw results data' in comparison to the raw data file produced during data acquisition (see 'data file name'). The worksheets are called 'Height'.

Raw results data may need to be normalized to reduce the impact of between-series drifts of instrument sensitivity, caused by machine maintenance, aging and tuning parameters. Such normalization data sets are called 'norm data' worksheets.

The formula that is used for normalization is.

$$Metabolite_{Normalized} = \left({Metabolite_{Raw}}/{iSTD_{Sum}}\right) * iSTD_{Average\ sum}$$

For example, identical samples ('QC samples') that were analyzed multiple times in all series of data acquisitions can be used. In fact, one must not exclude the possibility that even within a series of data acquisitions, a sensitivity shift or drift might occur.

For that reason, the Fiehn laboratory uses a suitable QC sample for every 11th injection. Such QC samples need to be as similar to the actual biological specimen as possible, e.g. generated by pool samples during extractions or by obtaining typical community standard samples (e.g. the NIST standard blood plasma, or commercial serum or plasma samples as needed).

Relative Quantification ('semi-quantification'):

Data can be converted to 'relative quantifications', meaning they are normalized to the best suited internal standard for which we know the absolute concentration that we used in the spiking process. The best suited internal standard is defined as the internal standard that belongs to the same lipid class as the metabolite that needs to be normalized. For example, all phosphatidylcholine lipids are normalized to our internal standard PC (12:0/13:0). For annotated lipids that lack an internal standard for the class, we can normalize based on nearest internal standard by retention time. This is because chromatography roughly separates the different lipid classes in different retention time groups.

The benefit of relative quantifications is that these normalized values should be not dependent on between series drifts or shifts in machine sensitivity. The drawback, however, is that the quantification relies on the accuracy of the internal standard addition (pipetting), peak finding and the quantification of a single internal standard. Quantification errors of a single peak (internal standards) are necessarily larger than errors of sum parameters (like the mTIC values).

If the internal standards are used for relative quantifications, the following equation is used.

$$Concentation\ of\ Metabolite = \frac{\left({Metabolite_{Raw\ Peak\ Height}}/{iSTD_{Raw\ Peak\ Height}}\right) * Concentartion\ of\ iSTD}{Amount\ of\ Sample\ Extracted}$$

Internal Standard Concentrations in (3:10, v/v) = MeOH:MTBE Added During Extraction

Lipid Internal Standard	ng/mL	µmol	nmol
1_CE(22:1) iSTD	16919.3	23.9	23924
1_PE(17:0/17:0) iSTD	369.8	0.51	513.5
1_PG (17:0/17:0) iSTD	1479.1	1.97	1972
1_LPC(17:0) iSTD	246.5	0.48	483.7
1_Sphingosine(d17:1) iSTD	54.8	0.19	191.9
1_Ceramide (d18:1/17:0) iSTD	123.3	0.22	223.3
1_SM (d18:1/17:0) iSTD	98.6	0.14	137.5
1_FA (16:0)-d3 iSTD	171.2	0.66	662.4
1_PC(12:0/13:0) iSTD	9.9	0.015	15.5
1_Cholesterol d7 iSTD	493.0	1.25	1252.3
1_TAG d5(17:0/17:1/17:0) iSTD	154.1	0.18	180.8
1_DG(12:0/12:0/0:0) iSTD	493.0	1.08	1079.6
1_DG(18:1/2:0/0:0) iSTD	2958.2	7.42	7422.1
1_MG (17:0/0:0/0:0) iSTD	986.1	2.86	2862.1
1_LPE (17:1) iSTD	123.3	0.26	264.8