Data Dictionary Fiehn laboratory _NIH West Coast Metabolomics Center_09-27-2013

Lipidomics by CSH-ESI QTOF MS/MS

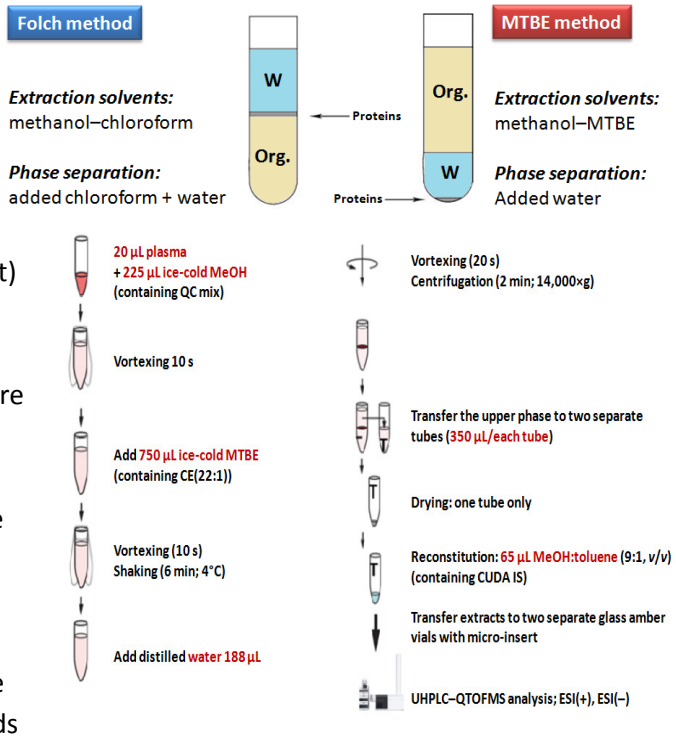
Glossary

CSH charged surface hybrid column. Waters corporation, a reversed phase UHPLC column.
UHPLC ultra high pressure liquid chromatography
ESI electrospray ionization. The method uses both negative ESI and positive ESI for negatively charged and 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 tandem mass spectrometry. After soft ionization by electrospray, the precursor (intact) charged molecules are fragmented 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
CE cholesteryl esters
PC phosphatidyl cholines (LPC is lyso-PC, see below)
PE phosphatidyl ethanolamines (LPE is lyso-PE, see below)
PS phosphatidyl serines
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)
IS internal standards
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

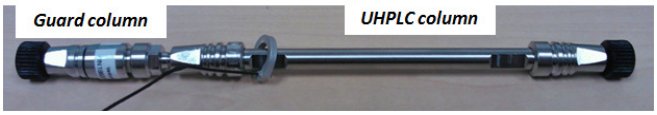
Extraction

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 on the right) 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

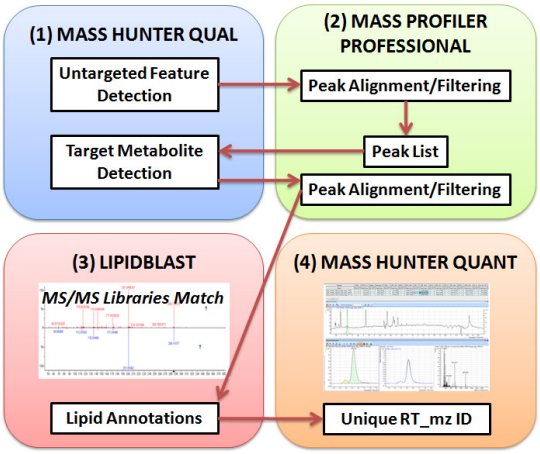
Data are acquired using the following chromatographic parameters:
Column: Waters Acquity UPLC CSH C18 (100 mm length x 2.1 mm internal diameter; 1.7 µm particles)
Mobile phase A: 60:40 v/v acetonitrile:water + 10 mM ammonium formiate + 0.1% formic acid
Mobile phase B: 90:10 v/v isopropanol:acetonitrile + 10 mM ammonium formiate + 0.1% formic acid
Column temperature: 65°C
Flow-rate: 0.6 mL/min
Injection volume: 1.67 µL for ESI(+) and 5 µL for ESI(–)
Injection temperature: 4°C
Gradient: 0 min 15% (B), 0–2 min 30% (B), 2–2.5 min 48% (B), 2.5–11 min 82% (B), 11–11.5 min 99% (B), 11.5–12 min 99% (B), 12–12.1 min 15% (B), 12.1–15 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(+); 40 eV for ESI(–)
Scan range: m/z 60–1700 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 column (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 (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

Data are analyzed in a four-stage process.
First, raw data are processed in an untargeted (qualitative) manner by Agilent's software MassHunter Qual v. B.05.00 to find peaks in up to 300 chromatograms. Peak features are then imported into MassProfilerProfessional for peak alignments to seek which peaks are present in multiple chromatograms, using exclusion criteria by the minimum percentage of chromatograms in which these peaks are positively detected. We usually use 30% as minimum criterion. In a tedious manual process, these peaks are then collated and constrained within the MassHunter quantification software v. B.05.01 on the accurate mass precursor ion level, using the MS/MS information and the LipidBlast library to identify lipids with manual confirmation of adduct ions and spectral scoring accuracy. MassHunter Quan enables back-filling of quantifications for peaks that were missed in the primary peak finding process, hence yielding data sets without missing values. The procedure is given in the panel to the left as workflow diagram.



Peaks are identified 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).

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. This identifier is set for a given method and does not change over time. It is given for both identified and unidentified metabolites in the same manner.
The **'name'** denotes the name of the metabolite, if the peak has been identified. A chemical name is not a unique identifier. We use names recognized by biologists instead of IUPAC nomenclature.
If a compound is identified, it has a name, and external database identifiers such as InChI key and LIPIDMAPS ID. If a compound is unknown, the name is the same as given in the 'identifier column'.
The **'elemental composition'** denotes the formula of the metabolite, if the peak has been identified.

The **'comment'** denotes comments. Most regularly, we use the comment field to clarify which ion species (metabolite charged adduct) was used for quantification.
The **'LIPIDMAPS'** identifier gives the unique identifier associated with an identified lipid in the community database LIPIDMAPS.
The **'InChI key'** identifier gives the unique chemical identifier defined by the IUPAC and NIST consortia.
The **'internal standard'** column clarifies if a specific metabolite has been added into the extraction solvent as internal standard. These internal standards serve as retention time alignment markers, for quality control purposes and for absolute quantifications.
The **'batch mz'** column details the m/z value that was detected in a specific data processing sequence of chromatograms. This value may be slightly different from the mz value given in the 'identifier column'.
The **'batch rt'** column details the retention time that was detected in a specific data processing sequence of chromatograms. This value may be slightly different from the rt value given in the 'identifier column'.

								Subject ID	223913	157819	124940
								Local code	A0118502A	A0115659A	A0133456A
								Vial Barcode	1RAR7	1GZR9	1AN1N
								Date receivec	4-Dec-12	4-Dec-12	4-Dec-12
								Date of evalu	14-Feb-13	14-Feb-13	14-Feb-13
								Sample Status			
								REVISION			
								Comments	CSH_posESI_Q	CSH_posESI_Q	CSH_posESI_Q
								Acq. Date-Tir	#########	#########	#########
								Data File Nam	B1_SA0001_TI	B1_SA0002_TI	B1_SA0003_TI
identifier	name	formula	comment	LipidMAPS	InChI key	internal stan	batch_mz	batch_rt			
0.78_341.28	CUDA	C19H36N2O3	[M+H]+	no entry	HPTJABJPZM	istd	341.2799	0.78	209276	193114	208345
1.04_286.28	Sphingosine (	C17H35NO2	[M+H]+	LMSP010400	RBEJCQPFFC	istd	286.2752	1.05	107806	86635	87168
1.34_466.29	LPE 17:1	C22H44NO7I	[M+H]+	LMGP02050(	LNJNONCNAS	istd	466.2925	1.35	69154	55579	55731
1.82_510.36	LPC 17:0	C25H52NO7I	[M+H]+	LMGP01050(	SRRQPVVYXI	istd	510.3551	1.84	707687	577154	547244
3.03_345.30	MG 17:0/0:0,	C20H40O4	[M+H]+	LMGL01010(	SVUQHVRAG	istd	345.2999	3.09	74532	67939	64973
3.17_421.29	DG 18:1/2:0/	C23H42O5	[M+NH4]+	no entry	PWTCCMJTP	istd	421.2925	3.23	464696	434328	421719
3.49_636.46	PC 12:0/13:0	C33H66NO8I	[M+H]+	LMGP01010(	FCTBVSCBB\	istd	636.4596	3.56	66807	55981	55165
4.26_479.37	DG 12:0/12:(	C27H52O5	[M+NH4]+	LMGL02010:	OQQOAWVK'	istd	479.3707	4.36	123139	116038	119977
4.81_376.40	Cholesterol d	C27H39D7O	[M-H2O+H]+	LMST01010(	HVYWMOML	istd	376.3955	4.93	61179	52093	59266
5.06_717.59	SM 17:0	C40H81N2O(	[M+H]+	LMSP030100	YMQZQHIES(	istd	717.5914	5.19	180663	131169	154248
5.95_552.54	Ceramide C1:	C35H69NO3	[M+H]+	LMSP020100	ICWGMOFDL	istd	552.5350	6.11	212902	175209	161458
6.23_720.56	PE 17:0/17:0	C39H78NO8I	[M+H]+	LMGP02011:	YSFFAUPDXI	istd	720.5561	6.41	308869	247214	234193
10.98_869.83	TG d5 17:0/1	C54H97D5O(	[M+NH4]+	LMGL03010(	OWYYELCHN	istd	869.8329	11.17	358760	286446	222391
11.71_724.70	CE 22:1	C49H86O2	[M+NH4]+	LMST01020(	SQHUGNAFK	istd	724.6966	11.86	142338	123855	118649
1.23_520.34	LPC 18:2	C26H50NO7I	[M+H]+	LMGP01050(	SPJFYYJXNPE		520.3395	1.24	450668	998939	449344
1.47_496.34	LPC 16:0	C24H50NO7I	[M+H]+	LMGP01050(	ASWBNKHCZ		496.3395	1.48	3836583	1863185	1402174
1.62_522.36	LPC 18:1	C26H52NO7I	[M+H]+	LMGP01050(	YAMUFBLW(		522.3551	1.63	581444	377007	279551
1.79_548.37	LPC 20:2	C28H54NO7I	[M+H]+	LMGP01050:	YYQVCMMXF		548.3708	1.81	7746	5023	3730
2.23_524.37	LPC 18:0	C26H54NO7I	[M+H]+	LMGP01050(	IHNKQIMGV.		524.3708	2.26	1512584	654412	530437
2.37_550.39	LPC 20:1	C28H56NO7I	[M+H]+	LMGP01050(	GJTDRNFWII		550.3864	2.41	15438	5691	5096
3.05_552.40	LPC 20:0	C28H58NO7I	[M+H]+	LMGP01050(	UATOAILWG		552.4021	3.11	19133	8781	6632

Row metadata that are requested by a specific consortium are labeled in blue. Those rows may be different dependent on the information the Metabolomics laboratory obtains from a specific consortium.
Consortium **'subject ID', 'local ID', 'vial barcode'** detail information given by the consortium.
The row '**date received'** is the date when samples were received in the metabolomics laboratory.
The row '**date of evaluation'** is the data of data acquisition with time stamp, as given by the machine logbook.
The row '**sample status'** uses the consortium sample status code when samples have errors. The consortium sample status code does not give a code when data acquisition occurred without problems.
The row '**revision'** details if data processing yields a new data sheet. Data revisions may be needed when new algorithms have been tested, validated and deployed that might yield better raw data analyses than prior submissions. By default, therefore, data revisions replace the (less valid) prior data submissions. However, data revisions may also indicate a different form of data treatment, e.g. data normalizations (see below). In this case, the 'revision' would indicate the type of normalization.

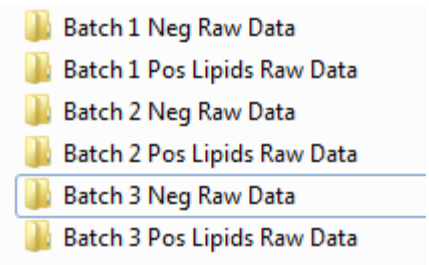
Any information in the row 'revision' will have a date stamp when the revision was conducted in the form of *MMDDYY*.
The **'comments'** row gives comments about the platform and type of sample. A sample is given as "sample" in comparison to e.g. a quality control or a blank injection.
The '**Acq.Date-Time'** row details the acquisition date and time when the data acquisition was completed.
The **'Data File Name'** row denotes the name of the raw data file. Raw data files are secured at the NIH Metabolomics database, www.metabolomicsworkbench.org
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
In case a sample will need to be reinjected, the file name will change from e.g.
B1_SA0001_consortiumnameLipids_Pos_1RAR7_.d to
B1_SA0001_consortiumnameLipids_Pos_1RAR7_2.d for the second injection,
B1_SA0001_consortiumnameLipids_Pos_1RAR7_3.d for the third injection and subsequent injections.
For shipment 1, raw file names with the ending _reinject were identical to the file ending _2. File names that comprise MSMS have MS/MS fragment spectra included (for compound identification); file names that do not comprise MSMS are for quantification only (on the MS level, without fragmentation). If a file further contains a _real acronym, this can be ignored (was used for laboratory internal QC measures).

Raw data files are submitted to the Data Coordinating Center in the following manner

a) **Filename**: Each file contains an SA number, this is the number represents the sample number in the order dictated by a specific consortium. Its primary function is for ease of sample tracking for the technician processing the LC samples. The purpose of prefixes such as B1_ is also only for laboratory organization. Neg or Pos represent the ESI ionization mode (+) or (-). The 5 digit number/letter combination represents e.g. a consortium sample identifier. An example of the file name is below. With respect to file name extensions, see above.

b) **Description of contents of each file**: Each file ending in .d is a raw Agilent data file, generated by the LC-QTOF MS instruments. Files that have SA in the file name are samples. Sample names containing QC## are injections of commercially available Citrate Plasma used as an in-run quality control. Samples that have NoInj or Blank in the name are controls for instrument carryover of samples (NoInj are runs without any injections, Blank injections inject extract resuspension buffer without samples to control for contamination of our resuspension buffers). Each folder also contains an additional Microsoft Excel spreadsheet that details the worklist used for sample acquisition.

c) **Number of Files**:



Each folder shown above may contain e.g. 300 sample files until we performed instrument maintenance work (e.g. change of guard columns, see the section *Data acquisition*), in addition to all QC and carryover control injections (NoInj and Blank) that were run together with these samples.

d) Format of file types: Raw data files are ".d" files generated by Agilent QTOF mass spectromters. They can be viewed using Agilent software such as MassHunter Qualitative software.

e) Extension of file types: See above under 'file names'.

The **result data sheets** 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 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.

There are many different types of normalizations in the scientific literature. We usually provide first a variant of a 'vector normalization' in which we calculate the sum of all peak heights for all identified metabolites (but not the unknowns!) for each sample. We call such peak-sums "mTIC" in analogy to the term TIC used in mass spectrometry (for 'total ion chromatogram/current'), but with the notification "mTIC" to indicate that we only use genuine metabolites (identified compounds) in order to avoid using potential non-biological artifacts for the biological normalizations, such as column bleed, plasticizers or other contaminants.
Subsequently, we determine if the mTIC averages are significantly different between treatment groups or cohorts. If these averages indeed are different by $p<0.05$, data will be normalized to the average mTIC of each group. If averages between treatment groups or cohorts are not different, or if treatment relations to groups are kept blinded, data will be normalized to the total average mTIC.

Following equation is then used for normalizations for metabolite *i* of sample *j*:

$$\text{metabolite}_{ij,\,\text{normalized}} = \frac{\text{metabolite}_{ij,\,\text{raw}}}{\text{mTIC}_j} \cdot \text{mTIC}_{\text{average}}$$

The worksheet is then called **'norm mTIC'**. Data are 'relative semi-quantifications', meaning they are normalized peak heights. Because the average mTIC will be different between series of analyses that are weeks or months apart (due to differences in machine sensitivity, tuning, maintenance status and other parameters), **additional normalizations** need to be performed. For this purpose, identical samples ('QC samples') must be analyzed multiple times in all series of data acquisitions. In fact, one must not exclude the possibility that even within a series of data acquisitions, a sensitivity shift or drift might occur. Hence, the following statistical analyses are suggested: (a) compute univariate statistics for mTIC values in batches within-series and between-series of data injections, using time/date stamps to find potential breaks during which machine downtime may have occurred. If there are no mTIC differences between such time/date stamp batches, calculate an overall mTIC covering all samples. (b) compute multivariate PCA plots for the , marking the potentially different samples of individual time/date stamp batches using different colors. If there is no apparent separation between PCA clusters of different colors, there is no

large between-series effect and these PCA clusters can be treated as indistinguishable. If there is suspicion of hidden features that might be masked by overall variance analysis in PCA, supervised statistics by Partial Least Square regression models can unravel such between-series differences.
Once different clusters (i.e. series of undistinguishable QC samples) have been identified, correction factor models need to be developed that correct differences between those QC samples. Subsequently, these correction factors can be applied to the actual analytical samples to remove overt quantification differences that are not related to biological causes but solely due to analytical errors.
Such correction factor models can be computed in different ways, e.g. by unit-variance mean centering or by calculating simple offset vectors for each individual metabolite. The best way of such types of normalizations is being explored in the Fiehn laboratory. However, in any case, such correction models can only be developed if a sufficient number of QC samples have been included in the analytical sequences. 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).

If the internal standards are used for absolute quantifications, the following equation is used for peak height normalizations for metabolite *i* of sample *j* and internal standard *k*

$$\text{metabolite}_{ij,\,\text{normalized}} = \frac{\text{metabolite}_{ij,\,\text{raw}}}{\text{istd}_k} \cdot \text{concentration istd}_k$$

The worksheet is then called **'norm istd'**. Data are 'absolute 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 unidentified lipids, we do not know the exact lipid class. However, because chromatography roughly separates the different lipid classes in different retention time groups, we can use the closest eluting internal standard for normalizing unidentified metabolites in order to get a rough estimate of a likely absolute concentration.
The benefit of absolute 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). We are currently evaluating the benefits or disadvantages between both types of normalization strategies (**norm mTIC** versus **norm istd**). This evaluation has not been completed yet in the Fiehn laboratory.