Data Dictionary Fiehn laboratory _NIH West Coast Metabolomics Center_05-2014

Analysis of special metabolites by HILIC-ESI QTOF MS/MS

Glossary

HILIC hydrophilic interaction chromatography, a variant of normal-phase chromatography

UHPLC ultra high pressure liquid chromatography

ESI electrospray. The method uses both negative ESI and positive ESI for negatively charged and positively charged molecules.

QTOF quadrupole time of flight mass spectrometer

MS/MS tandem mass spectrometry. After soft ionization by electrospray, the precursor (intact) charged molecules are fragmented by collision with gas atoms, usually Helium. 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.

MTBE methyl-tertiary butyl ether

MeOH methanol

BEH amide bridged ethylene hybrid amide column

QC quality control

IS, istd internal standards

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.

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.

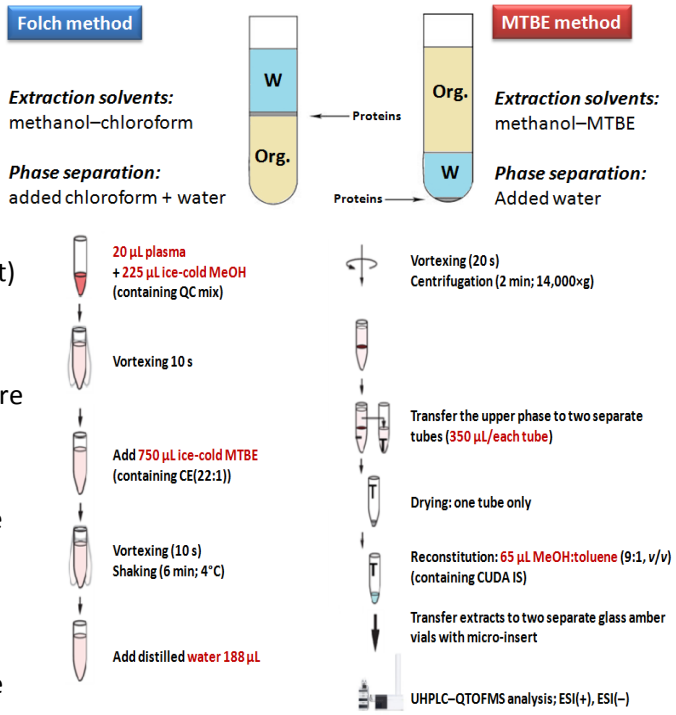
IUPAC International Union of Pure and Applied Chemists

NIST National Institute of Standards and Technology

PCA Principal Component Analysis

Extraction

Blood plasma or serum is best 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 are continuing to optimize choices of internal standards, e.g. labeled TMAO and similar compounds. The top layer is used for lipidomics while the bottom layer (methanol/water phase) is very suitable for HILIC-MS investigations.



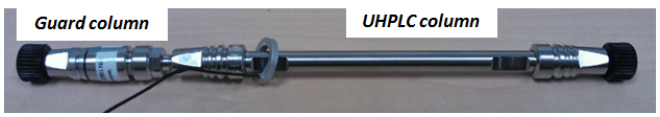
Alternatively, other extraction methods can be used, for example, a mixture of acetonitrile/water/isopropanol (2:2:3). In such cases, complex lipids would be found in HILIC-QTOF MS chromatograms in addition to special metabolites such as betaine, choline and TMAO.

Table . Chromatographic parameters for HILIC- QTOF MS/MS (polar compounds)

Column	Waters Acquity UPLC BEH Amide Column, 1.7 µm, 2.1 mm x 150 mm) Pre-column: Waters Acquity UPLC BEH Amide VanGuard pre-column (1.7 µm , 5 mm × 2.1 mm;)
Mobile Phase A	Ultrapure water with 10 mM ammonium formiate + 0.125% formic acid, pH 3
Mobile Phase B	95:5 v/v acetonitrile:ultrapure water w/ 10 mM ammonium formiate + 0.125% formic acid, pH 3
Column temperature	40°C
Flow-rate	0.4 mL/min
Injection volume	3 µL for ESI (+)
Injection temperature	4°C
Gradient	0 min 100% (B), 0-2 min 100% (B), 2-7 min 70% (B), 7.7-9 min 40% (B), 9.5-10.25 min 30% (B), 10.25-12.75 min 100% (B), 16.75 min 100% (B)
ESI capillary voltage:	+4.5 kV for ESI (+)
Precursor isolation width	3 Da
Collision energy	+45 eV for ESI (+)
Scan range	m/z 60-1200 Da
Spectral acquisition speed	2 spectra/sec
Mass resolution	10,000 for ESI (+) on an Agilent 6530 QTOF MS

Data acquisition

Data are acquired using the following chromatographic parameters, see table. 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 polar compound retention times or reproducibility of quantifications. This chromatography method yields excellent retention and separation of



metabolite classes (biogenic amines, cationic compounds) with narrow peak widths of 5–20 s and very good within-series retention time reproducibility of better than 6 s absolute deviation of retention times.

Data processing

Data are analyzed in a four-stage process.
First, raw data are processed in an untargeted (qualitative) manner by the free mzMine 2.0 software l to find peaks in up to 300 chromatograms. Alternatively, selected peaks can be collated and constrained into Agilent's MassHunter quantification method on the accurate mass precursor ion level, using the MS/MS information and the NIST14 / Metlin / MassBank libraries to identify metabolites with manual confirmation of adduct ions and spectral scoring accuracy. MassHunter enables back-filling of quantifications for peaks that were missed in the primary peak finding process, hence yielding data sets without missing values.

Data reporting

Data are reported including metadata.
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.
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 **'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'.
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

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 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 type 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'), 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*:

$$\text{metabolite}_{ij,\text{normalized}} = \text{metabolite}_{ij,\text{raw}} / \text{mTIC}_{j} * \text{mTIC}_{\text{average}}$$

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 are 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 11^{th} 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).