Article

Solutions for low and high accuracy mass spectrometric data matching. A data-driven annotation strategy in non-targeted metabolomics

Sara Forcisi, Franco Moritz, Marianna Lucio, Rainer Lehmann, Norbert Stefan, and Philippe Schmitt-Kopplin

Anal. Chem., **Just Accepted Manuscript** • DOI: 10.1021/acs.analchem.5b02049 • Publication Date (Web): 21 Jul 2015

Solutions for low and high accuracy mass spectrometric data matching. A data-driven annotation strategy in non-targeted metabolomics

Sara Forcisi,[§,♯,∞,] Franco Moritz,[§,♯,∞] Marianna Lucio,[♯] Rainer Lehmann [‡,¶,∞] Norbert Stefan,[¶, ∞,≈] and Philippe Schmitt-Kopplin [♯,★,∞,*]*

§These authors contributed equally.

[♯] Research Unit Analytical BioGeoChemistry, Helmholtz Zentrum München, German Research Center for Environment Health, Neuherberg, D-85764, Germany

[★] Chair of Analytical Food Chemistry, Technische Universität München, Freising-Weihenstephan, D- 85354, Germany

[‡] Division of Clinical Chemistry and Pathobiochemistry (Central Laboratory), University Hospital Tübingen, D-72076, Germany

[¶] Institute for Diabetes Research and Metabolic Diseases of the Helmholtz Centre Munich at the University of Tübingen (IDM), D-72076 Tübingen, Germany

[∞] German Center for Diabetes Research (DZD), Germany

[≈] Department of Internal Medicine, Division of Endocrinology, Diabetology, Nephrology, Vascular Disease and Clinical Chemistry, University Hospital of the Eberhard Karls University, D-72076 Tübingen, Germany

ABSTRACT: Ultra High Pressure Liquid Chromatography coupled to mass spectrometry (UHPLC-MS) has become a widespread analytical technique in metabolomics investigations, however the benefit of high performance chromatographic separation is often blunted due to insufficient mass spectrometric accuracy. A strategy that allows for the matching of UHPLC-MS data to highly accurate Direct Infusion Electrospray Ionization (DI-ESI) Fourier Transform Ion Cyclotron Resonance / Mass Spectrometry (FT-ICR/MS) data is developed in this manuscript. Mass difference network (MDiN) based annotation of FT-ICR/MS data and matching to unique UHPLC-MS peaks enables the consecutive annotation of the chromatographic dataset. A direct comparison of experimental *m/z* values provided no basis for the matching of both platforms. The matching of annotation-based exact neutral masses finally enabled the integration of platform specific multivariate statistical evaluations, minimizing the danger to compare artifacts generated on either platform. The approach was developed on a Non-Alcoholic Fatty Liver disease (NAFLD) dataset.

Metabolomics can be performed using a multitude of analytical instruments each with different sensitivities, specificities and frameworks for feature identification[1-4]. FT-ICR/MS offers high mass accuracy and resolution; the ion cyclotron resonance enables a highly sensitive, semi-quantitative detection of thousands of metabolites at the same time [5]. UHPLC-MS offers more quantitative mass spectrometric results with medium sensitivity and comparably low mass accuracy and resolution, along with the possibility to characterize features on the retention time level[5]. Within multi-platform approaches, comparability of the data produced by different techniques is crucial. Along these lines, data fusion can be used for the validation of findings and the improvement of feature identification. A multitude of papers suggest procedures for the fusion of LC-MS/GC-MS data [6,7], LC-MS/NMR data [8] or GC-MS/NMR data [9]. Fusion of DI-ESI-

FT-ICR/MS data to other metabolomics data is poorly described and complicated by the intrinsic semi-quantitative nature of this technique. Independent of their level, fusion techniques require a certain level of comparability in terms of quantification-capability and sources of artefacts. For example, if one technique is more strongly afflicted by matrix effects than the other technique (which is the case once direct infusion data and LC-MS data are compared) fusion approaches cannot work properly, since they are commonly based on correlation or covariance. Hence, feature matching and separate comparison of their statistics would be a more appropriate strategy. However, such a strategy requires high confidence sum formula annotations and most time of flight mass spectrometers used for UHPLC-MS do not deliver a high level of mass accuracy. High resolution/accuracy mass spectrometric scans require sampling rates < 1 Hz, for which reason they do not support ultra high chromatographic resolution [5]. Furthermore, most features that are detected via classical non-targeted metabolomics using full scan (UHPLC-)MS are too small for their isotopologues to be detected. The same is true for their tractability to tandem MS. Despite the advantageous retention time (RT) information of LC-MS, there are no useful RT-databases as the requirements of applied columns, gradients and solvents for different tasks and sample types are too divers to be covered by some proposed standard method. In consequence, a method that enables high confidence sum formula annotation for low accuracy LC-MS data, beyond the framework of classical methods that use available chemical standards, is required.

The aim of this work is the introduction of an effective solution to annotate low mass accuracy mass spectrometric data. The data are generated via DI-ESI-FT-ICR/MS (high mass accuracy) and Reversed-Phase (RP)-UHPLC-MS (low mass accuracy). In order to accomplish this task we resort to a mass difference network algorithm[10] based strategy for the annotation of the UHPLC-MS data. The first crucial step is the selection of high confidence starting points. They can be determined through multiple ways: combinatorial assignment (employing senior rules and isotopic patterns recognition), database matching, an internal standard database matching (when the samples are spiked prior analysis) and annotated data matching (when the annotated data derived

from a different platform acquisition). The latter case describes our choice, where high mass accuracy annotated data from DI-ESI-FT-ICR/MS allowed us to determine the starting points to generate a MiDN based on the experimental UHPLC-MS data. The second crucial step is the integration of the retention time of detected features into MDiN reconstruction to increase annotation performance for UHPLC-MS data. Here, both techniques are applied for the analysis of human blood plasma samples of individuals that suffer from Non-alcoholic fatty liver disease (NAFLD). NAFLD is a disease of our generation that belongs to the metabolic syndrome and it is related to obesity and diabetes mellitus. The rate of NAFLD is very high in industrialized countries; it was shown to involve 10-20% of the population [11-14]. The development and the progression of NAFLD can trigger cardiovascular disease (CVD) disorders and type 2 Diabetes, if they are not already affected by the latter. Fatty liver strongly correlates with insulin resistance, an important predictor of type 2 Diabetes and cardiovascular disease [11]. The condition of subjects that are affected by fatty liver disease without showing an insulin resistance pattern [15,16] is of high relevance, because this condition may support the identification of markers and mechanisms that are essential for insulin sensitivity. The proposed strategy enables high confidence sum formula assignment to low mass accuracy (UHPLC-)MS features, which is the fundamental prerequisite for a compound class-based comparison of multiple metabolomic platforms. Users of low resolution mass spectrometry in non-targeted metabolomics can initiate the workflow by either comparing their data to one high mass accuracy spectrum of a quality control or by any other means of high confidence formula assignment supported by careful isotopic pattern recognition.

EXPERIMENTAL SECTION

Metabotyping via FT-ICR/MS. The analyses of the samples were performed on a Brucker APEX Qe FT-ICR/MS with a TRiVersa Nanomate chip electrospray ionization system (Advion Ithaca, USA). The resolving power was 120,000 at *m/z* 400 in positive mode. The ionization was applied in positive mode within a mass range of 147 *m/z* to 2,000 *m/z* and with a time domain of 1 MW. 1,024 scans were acquired for each spectrum.

Metabotyping via UHPLC-QTOF/MS. LC-MS sample analyses were performed using a Waters Aquity UHPLC system coupled to a Synapt HDMS oa-Q-TOF mass spectrometer (Waters, Milford) equipped with an electrospray (ESI) ion source operating in positive mode. The chromatographic separation was performed by a C18 Vision HT-HL UHPLC column (2 x 150 mm, 1.5 µm, Alltech Grom GmbH, Germany). Elution buffer A was water containing 0.1% formic acid and elution buffer B was acetonitrile. The flow rate was set to 0.300 ml/min. The linear gradient method consisted in 5 % of B over 0-1.12 minutes, 5% to 100 % of B over 1.1-22.3 minutes and held at 100% B until 29.5 minutes, returned to 5% of B at 29.6 minutes. In order to equilibrate the column with the initial mobile phase, 5% B was kept until 35 minutes. The column oven was set to 40 ºC and the sample manager temperature to 4 ºC. A solution of leucine enkephalin (556.2771 *m/z*, 400 pg/µl) in MeOH/H_2O:1/1 containing 0.1 % of formic acid was infused as lockmass compound at a flow rate of 5 µl/min. The lock mass correction is applied on the experimental masses following the calibration curve generated during the instrument calibration. The samples were measured in random triplicates within three different batches, where one batch included all samples measured one time. The spectra were acquired in centroid mode within a range of 50-1000 *m/z*, the detection parameters are described in table S2.

Sample analysis was performed in the following order: 3 blanks at the beginning, 16 quality control plasma (QC) for column conditioning. A standard mix, diluted in 20 % acetonitrile was injected after QC plasma column conditioning. Subsequently, the three randomized batches were analyzed. Additional standard mixes inserted in the middle and the end of the entire study (each standard at the concentration of 1 mg/L). Quality control plasma (QC) were injected every 10^{th} measurement. Plasma blank solutions (i.e. water collected in plasma Sarstedt tubes and treated as plasma, see supplementary material) were randomly injected in order to examine a possible presence of carry-over and of contaminants coming from the sample preparation. Details on the standards chosen during the study are described in table S3 and table S4.

Treatment of FT-ICR/MS data. Mass spectra were externally calibrated on clusters of arginine

(1mg/ml in methanol /water: 80/20). The internal calibration was based on data base matching using the web server MassTRIX (http://masstrix.org) [17]. All the input masses were assigned to the closest reference mass found in the database, within 1ppm error. Consequently a new calibration list was created performing a regression through the error distribution of these MassTRIX annotations. The standard deviation of mass relative to the calibration masses within the *m/z* range of interest was always below 100 ppb. The calibrated spectra were exported at signal-to-noise-ratio of 4. The peaks of all exported spectra were aligned and stored in a matrix using in-house software at an error tolerance of ± 1 ppm. Data were then filtered for mass defects and screened for isotopic peaks and artifacts within an error range of 1 ppm. Peaks were removed if their Pearson correlation coefficients with their adjacent mono-isotopic peak were higher than 0.95. *M/z* values were then used for mass-difference network [MDiN] reconstruction. MDiNs consist of source nodes (*m/z* values or features of substrates) and target nodes (products) which are connected by edges (*m/z* differences that are equivalent to a compositional difference of a biochemical raction; Reaction Equivalent Mass Differences [REMDs]). Edges between nodes were assigned if the mass difference of two nodes matched a pre-selected REMD within an error range of ± 0.2 ppm. As MDiNs are likewise compositional difference networks, it is possible to assign a sum formula to any node that is available starting from a known node (with defined sum formula). After defining the sum formula of a starting node, the sum formulas of adjacent non-annotated nodes are updated by the combination of the information carried by the adjacent known node and the incident edge. After each iteration, all assigned sum formulas and edges are tested for consistency of the relationships between the source-formulas, the REMDs and the target-formulas. An assignment is removed if the information of 'source node' + 'edge' = 'target node' does not match. The result of the algorithm is a maximally self-consistent network. The final sum formula annotation error was set to ± 10 ppm, while the majority of annotations were found within an error range of ± 0.4 ppm and 0.22 ppm standard deviation; no assignment showed an error larger than 1 ppm.

Treatment of UHPLC-QTOF/MS data. External calibration was based on clusters of sodium formate. Internal calibration of the spectra was performed via lock mass correction, by orthogonal infusion of leucine enkephaline. The data acquired through UHPLC-QTOF/MS were aligned using MarkerLynx software (Waters, Milford) within a mass range of 0.02 Da and RT window of 0.1 minutes. The chromatograms acquired were processed using ApexTrack peak integration to detect chromatographic peaks. The masses were aligned and normalized with total area peak normalization to the sum of 10,000.

Data were then treated in the same way as the FT-ICR/MS data. Edge assignment errors were set to ± 0.5 ppm and the final annotation error tolerance was set to ± 20 ppm. The majority of annotations self-assembled within a range of + 4 ppm and – 4 ppm.

Statistical evaluation. In order to reduce the impact of noise and to stabilize the variance among all the samples, various pre-processing steps have been applied. The total frequency threshold of *m/z* value occurrence across the samples was set to 30%. The variable's zero values were counted across all samples and the signals appearing in less than 30% were excluded from further analysis To disclose valuable biological information from the data, several multivariate analyses were performed. In addition, different visualization tools were applied for the interpretation of the results. Therefore, our analysis involves a sets of software such as SIMCA-P+12 (Umetrics, Umea, Sweden) and MATLAB R 2011 (The MathWorks, Inc., Natick, Massachusetts, United States). Several models were built and validated in order to reduce the data dimensionality (Principal Component Analysis) and to retrieve the discriminant metabolites via OPLS (Orthogonal PLS modeling). The dataset was UV scaled for UHPLC-MS data and ctr (centered but not scaled) for FT-ICR/MS data. Seven-fold cross validation was used to assess the internal validity. Different model parameters have been evaluated in order to assess the goodness of the model: R^2 and Q^2 which estimate the goodness of fit and prediction, respectively. Moreover, to determine the reliability of the OPLS model, the diagnostic tool CV-ANOVA (ANalysis Of VAriance testing of Cross-Validated predictive residuals) has been performed.

Data base annotation. All theoretical neutral masses or formulas of annotated data were matched to HMDB 3.5 (http://www.hmdb.ca/) [18] for compound assignment.

Detailed information on used chemicals, patients, plasma sample collection and treatment as well as details about the instrumental analytics are described in the supplementary material.

RESULTS AND DISCUSSION

Matching strategy preface. As elegantly shown by Vaughan et al.[19], the comparison of different UHPLC-MS methods with similar column chemistry is facilitated by the availability of retention time data for the sets to be compared. The same is not possible when UHPLC-MS data needs to be compared to DI-ESI-FT-ICR/MS data, which have ultra-high mass accuracy, resolution and sensitivity but no time dimension. MS hyphenation to UHPLC can achieve isobaric/isomeric separation; however the lower mass accuracy and resolution of TOF/MS does not immediately support the differentiation between isobaric and isomeric patterns if isotopologue peaks of one compound merge with the monoisotopic peak of another compound. For this reason, database matching of LC-TOF/MS features prior to accurate data annotation has to be taken with care. On the other hand, due to the absence of a time dimension and due to the semi-quantitative nature of FT-ICR/MS, data comparison with UHPLC-MS can only be performed on the *m/z* dimension. It is erroneous to assume that *m/z* data is free of error after calibration, because calibration merely removes systematic error. The remaining random error commonly has a spread of ±0.2 ppm in FT-ICR/MS data, and ±1 to ±5 ppm (or more) are typical errors for TOF/MS. Often, error distributions are not centered around 0 ppm, which makes comparison even more difficult.

The investigation of the optimal matching strategy started with the assessment of the FT-ICR/MS data acquired in positive ionization mode, where 17,934 ions were detected. After a 10% cut off in frequency of masses over samples sum formula assignment was conducted on 9,442 masses. MDiN-based annotation using 176 REMDs resulted in 2855 (30%) mono-isotopic sum formulas. From these putatively assigned formulas, 859 were found in HMDB 3.5.

Using UHPLC-MS in positive ionization mode 13,268 ion features were recorded on the same

plasma samples. Features were accepted if at least 3 out of 39 triplicates were populated with at least 2 non-zero values. After this cut off, the assignment and the statistical analysis were performed on 11,639 features. The first MDiN reconstructed using 176 REMDs encompassed 25,355 edges over 7,606 nodes.

Low mass accuracy UHPLC-MS annotation workflow. In order to understand whether the relative *m/z* peak positions between both datasets have a functional relation, both data sets were matched using a ± 10 ppm error window and the mass deviation between the matched pairs was plotted over their *m/z* values (Figure S1). No reasonable error distribution could be approximated when the entire sets of experimental *m/z* features were matched (Figure S1). Any variation of the matching error window would not have improved this result. Matching the 2,855 theoretical *m/z* values of the FT-ICR/MS-set against the UHPLC-MS features revealed a more detailed error over *m/z* distribution at ±4 ppm (Figure 1A). Overall, 1,616 FT-LC feature-pairs were found within the ± 10 ppm error window. 306 FT-MS features and 1,438 LC-MS features were found to be unique (not occurring multiply), and 242 feature pairs were found to be biunique. 136 pairs were located within the observed ± 4 ppm error distribution and 125 paired LC-MS features were part of the above created FT-ICR/MS-MDiN. For reasons of resolution dependent annotation impairment, we decided to use MDiN-based sum formula assignment on the UHPLC-MS dataset. Figure 1A) highlights the reasoning behind the appropriate selection of starter masses for MDiN-based LC-MS data annotation. If FT-ICR/MS and UHPLC-MS features are not connected in a biunique manner, it is more probable for isobaric UHPLC-MS features to be annotated as being of isomeric nature. Consequently, already the initiation of the MDiN-based LC-MS data annotation would be leveraged towards inappropriate isobaric spaces.

MDiN-based annotation was initiated with the formulas of the 125 biunique LC-MS – FT-MS pairs. The ±0.5 ppm error window that was applied for network reconstruction leaves many degrees of freedom for the false positive assignment of REMDs. It was therefore to be expected, that MDiN-annotation would converge to several locally optimal solutions. In

consequence, 100 repetitions of MDiN-annotation were performed and compared.

The next step illustrated in Figure 1B) highlights the frequency distribution of isobaric annotation counts per feature and the obtained error over *m/z* distribution of the stably annotated features. In order to underline the goodness of this result, the LC-MS data set was annotated with a combinatorial in-house written software. Within a ±2 ppm error window 572,689 Senior-valid formulas were found. The average amount of isobaric annotations per feature was calculated to be 109.8 and only 259 features received a unique annotation (Figure S2). Nonetheless, 4,574 out of 5,564 MDiN-annotated features (81.7%) remained inappropriate for a comparison of annotation based theoretical LC-MS features with the theoretical FT-ICR/MS data. Especially features of the higher mass range received multiple isobaric annotations. To improve this result RT-information needed to be included into the annotation procedure.

Recently, Morreel et al.[20] proposed to use candidate substrate-product pairs (CSPP), which are equivalent to the REMDs in this manuscript. CSPPs are supposed to be used only if the sign of the RT-difference between two *m/z* features is the same as an priori specified CSPP-RT shift. Screening the stable UHPLC-MS annotations of the present NAFLD data set for REMDs of 100% specificity for RT shifts (positive or negative exclusively), revealed that REMDs (which carry a hypothesis in regard to their chemical functionality) are inappropriate for the performance of a RT supported MDiN-annotation. Hexadecanoic acid condensation was found to occur 110 times with increasing RT (RT^+) and 8 times with decreasing RT (RT^-). Considering the amount of possible isobaric annotations, nothing else than a proportion of RT^+:RT^- = X:0 or 0:Y is to be accepted for an RT-supported network annotation. Only 3 REMDs were found in proportions of X>20:0 and none was found to be uniquely RT^-specific. Only 12 out of 176 REMDs were found to be specific at all. As the assignment of RT shifts to REMDs did not match pre-specified expectations (e.g. condensation with Glucose yielded: RT^+: 21; RT^-: 36), all stably annotated features were screened for any positive mass differences that showed 100% RT-specificity (Figure 1C)). The screening resulted in 876 RT-specific mass differences (RTMDs) for

RT^+ and 162 RTMDs for RT^- (observed for at least 20 pairs). The maximum observed RTMD frequency was 106 for RT^+ and 71 for RT^-. The formation of sodium adduct ions was defined to occur at the same RT as their protonated equivalents.

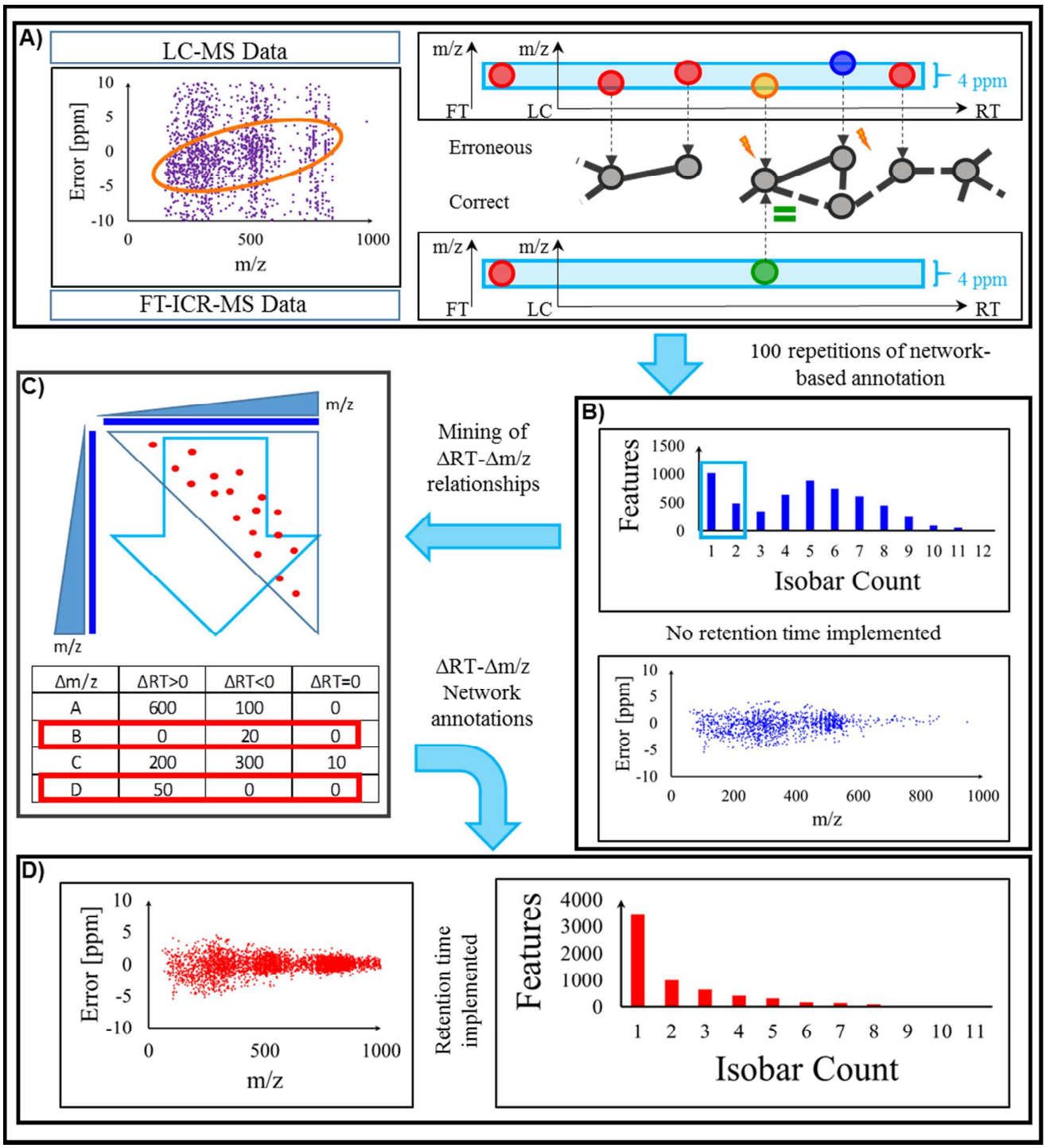


Figure 1. Convolution of the RT-supported UHPLC-TOF-MS annotation strategy. A) Error over *m/z* plot of theoretical FT-ICR/MS ion masses to UHPLC-TOF-MS data matching and reasoning behind the selection of

biunique Hits for MDiN annotation. The isomeric annotation of possible isobars may project multiple false starting points into the MDiN. B) Summary of 100 repetitive MDiN annotations. More than 80% of all features obtain multiple isobaric annotations. The error over *m/z* distribution of the stable annotations is well centered. Features with *m/z* > 500 are poorly annotated. C) The stable annotations, their RTs and Formulas are screened for RT-specific mass differences. Only mass differences with absolute RT+/--specificity (at least 20 vs. zero observations) are considered for MDiN reconstruction. D) Summary of 100 repetitive RT-MDiN annotations. The number of stable annotations is increased more than 3-fold; the proportion of unstable annotation is decreased. The corresponding error over *m/z* distribution is well centered and features with *m/z* > 500 are well addressed.

The 1,038 found RTMDs were used to reconstruct an RT-directed MDiN and their corresponding formula differences were used for feature annotation. As shown in Figure 1D), 100 repetitive annotations revealed a 3-fold increase of unique isobaric annotations and a decrease in the count of features with multiple isobaric annotations. Ultimately, 4,564 annotations which had at maximum 3 isobaric assignments of which the most abundant isobar was found in at least 66 out of 100 repetitions were chosen to be transferred to the LC-MS data set.

The initial matching of theoretical FT-ICR/MS data contained 1,616 feature pairs among which 984 pairs were found within a ± 4 ppm error window. After UHPLC-MS annotation, 424 of these pairs were confirmed. 57% of all pairs turned out to be false assignments. This result emphasizes the importance of careful feature annotation prior to any (database) matching. The authors use MDiN based annotation approaches, as they use the compositional context of a data set to navigate through different isobaric annotations. Naturally, any other means of reliable feature annotation can be used. Likewise, it is obligatory to validate assignments when *m/z*-matching to databases is used as a means of feature annotation.

Statistical analysis of FT-ICR/MS data. Data that went through the multivariate analysis pipeline were treated with a stricter feature frequency cutoff (30%). PCA analysis did not provide a good separation among the two classes (data not shown). Therefore, more sophisticated techniques were applied. An OPLS model (Figure

S3) could differentiate the two phenotypic classes. The model had one predictor component with ($R^2Y(cum)=0.5$) for the model fit, ($Q^2(cum)=0.4$) for the predictiveness in cross-validation and $p<0.05$ for CV-Anova. Outliers were not considered for further analysis. Covariance and correlation between the features and the model was inspected (not shown) by the S-Plot [21].

The final OPLS model contained 5,655 features, 1888 of which obtained a CHNOPS formula via MDiN annotation. The annotated features within the OPLS model were queried against the HMDB 3.5 database in order to assess their probable compound classes. 771 sum formulas (40.8%) were successfully matched against the HMDB database. In the spirit of gene set enrichment analysis [22], it was of interest to gain insights upon each compound classes' general location on the discriminating latent variable. In order to avoid the dominant behavior of single features with extreme weights, the variables were ranked from 1 to N and the ranks were then centered according to the central rank of the weights. Afterwards, the sum of ranks of each compound class with more than 5 observations was determined. Positive rank sums indicated compound class overrepresentation in IS and negative average ranks indicated otherwise. The rank sums were then normalized on the maximum norm (Figure S4).

Statistical analysis of UHPLC-QTOF/MS data. The UHPLC-QTOF/MS data set was evaluated via multivariate statistics [23,24] in order to obtain insights on the data. Partial least squares discriminant analysis (PLS-DA) was performed on the two classes of observations: IR (class1) and IS (class2). The model was unable to separate the two groups of observations. OPLS modeling was applied in order to define features with discriminative power. The model consisted of one predictor component, and two orthogonal components. The generated model gave a good value for the model fit ($R^2Y(cum)=0.98$) and for the predictiveness in cross-validation ($Q2(cum)=0.4$), with the CV-Anova $p<0.001$ (Figure S5). The detected outliers were not considered for further analysis. Centered ranks were calculated based on the predictor component (Figure S6).

The final model encompassed 4,048 variables, 1,965 of which could be annotated with a CHNOPS formula. The sum formula annotation of each feature was then queried against HMDB,

which resulted in a successful compound class assignment of 485 features (293 sum formulas). Positive maximum normalized rank sums indicated a tendency for IS-specific behavior, negative rank sums indicated otherwise.

Figure S7 and Figure S8 describe the classes of compounds which are up- and down-regulated in the insulin sensitive class, respectively. The different classes of metabolites are depicted as function of retention time (RT). Sizes of the single dots in Figures S7 and S8 are proportional to the amount of isomers stored in HMDB. The class of glycerophosphocholines is the most abundant among the up-regulated metabolites that separated at the end of the RP-chromatographic gradient. These results confirm the findings of the published targeted analysis of the same sample set[16]. Among the down- regulated features, more hydrophilic classes of compounds such as monosaccharides, glycosyl-compounds and amino acids occur during the first part of the RP-gradient.

Data matching. Different analytical techniques (e.g. DI-FT-ICR/MS and UHPLC-TOF/MS) provide data of inherently different quality. The best FT-ICR/MS data quality can be achieved by superimposing multiple (hundreds of) scans acquired using DI-ESI. In this case, high sensitivity is accompanied by well defined, almost linear error over *m/z* distributions, which cannot be obtained via LC-FT-ICR/MS coupling. An appropriate description of UHPLC chromatographic peaks, with higher per-scan sensitivity, can be obtained via TOF/MS coupling. However, even if higher order polynomials are used for UHPLC-MS spectral calibration, the error distribution per mass can (i) not be guaranteed to be centered to zero at any time and at any *m/z* and (ii) the standard deviation of error distributions is unequally broader than that of FT-ICR/MS. Within the standard deviations of error distributions, each isobaric annotation is equally probable.

The problem that arises from multi-platform data matching is that the attempt to integrate such inherently different error distributions results in a multitude of equally probable matches, even if the systemic deviations in both data sets are perfectly calibrated (which cannot always be guaranteed in terms of the TOF/MS). Matching quality further impairs, when one (or both) of the peaks to be matched is (are) not centered at an error of 0 ppm.

As described above, both datasets were annotated by means of formula propagation through MDiNs [10], which reconstructs an MDiN over *m/z* data and performs formula assignment until all nodes in the network attain formulas that

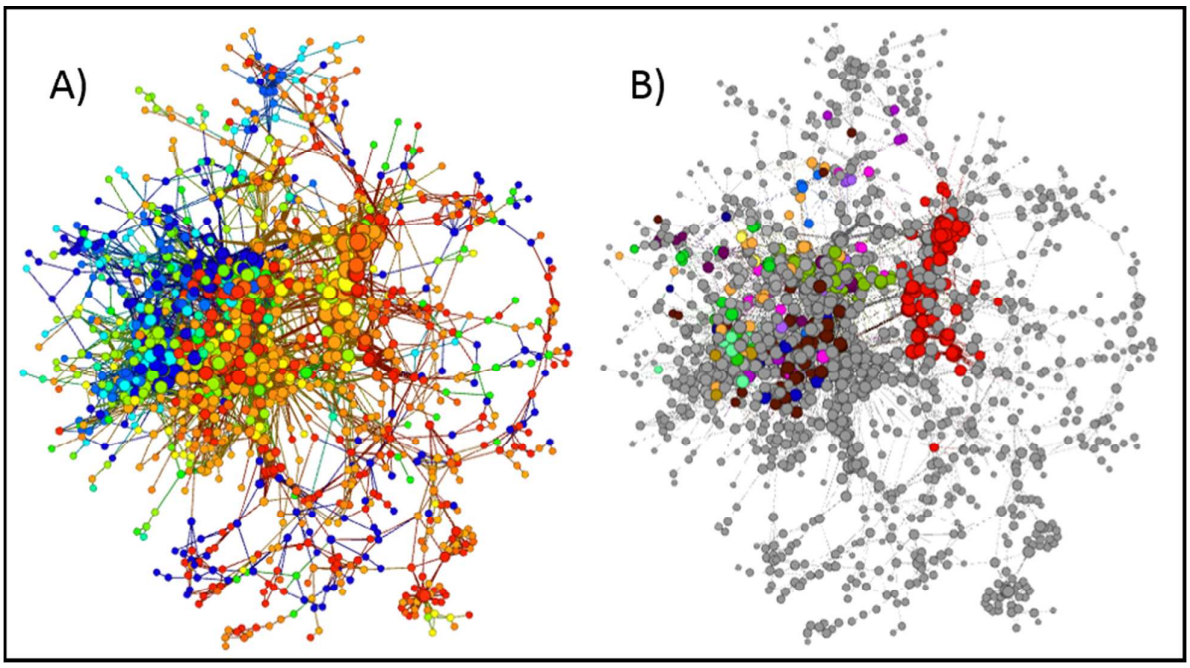


Figure 2. MDiN of annotated UHPLC-TOF-MS features in three different layouts: A) RT-gradient from blue over green to red. B) Glycerophospholipids (red), Prenol Lipids (brown), Carboxylic Acids (dark purple), Glycosyl Compounds (green), Amino Acids (orange), Fatty Alcohols (purple), Fatty Acids and Conjugates (pink), Fatty Acid Esters (olive), Fatty Amides (violet), Steroids and Derivateves (blue), Monosaccharides (yellow), Eicosanoids (light blue). Compound Classes with lower frequency or missing HMDB annotation are grey.

are consistent with their mass differences (and retention times).

To visualize whether HMDB compound class assignments associated with the reversed phase (RP) chromatographic regime, an MDiN over the statistical set of UHPLC-TOF/MS annotations was reconstructed using REMDs. Figure2A) shows a clear RT-gradient from blue to red (low RT to high RT) throughout the reconstructed MDiN. Figure 2B) shows the corresponding HMDB compound class assignments, which largely correspond to expectations given RP chemistry. The most prominent HMDB compound classes are highlighted in Figure 2B). Their

coloring follows the legends in Figures S7 and S8. While the amount of UHPLC-MS-HMDB formula matches was comparably low, known compound classes occur within clusters, which moreover match the RT-gradient in Figure 2A).

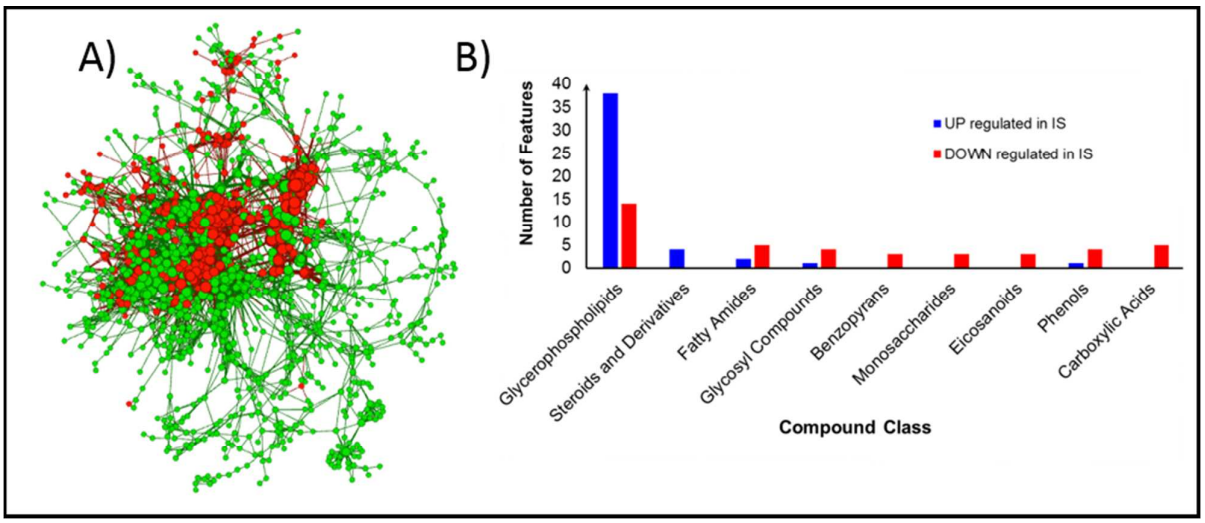


Figure 3. A) Positioning of corresponding FT-ICR/MS-UHPLC-TOF-MS features (red). B) Compound class distribution of the combined statistics. Up-regulated in IS is blue and down-regulated in IS is red.

Unknown classes and low abundance classes are colored in grey.

The coherence of both, the annotations and the statistical results was visualized by coloring the above obtained MDiN for the presence of RP-UHPLC-MS formulas in the DI-ESI-FT-ICR/MS dataset. The statistical feature ranks of both datasets were averaged and the corresponding compound class counts for correlation and inverse correlation with insulin sensitivity are shown as a bar chart (Figure 3).

The red nodes in Figure 3A) mark neutral formulas that were found in FT-ICR/MS. A glance at Figure 2B) reveals that the most frequently annotated network regions are populated by glycerophospholipids, prenol lipids and fatty acid esters. The remainder is covered by amino acids, glycosides and other hydrophilic classes. The compounds which share the same formula as the FT-ICR/MS features are concentrated at the center of the network and can therefore be taken to be major compounds of human plasma. The outer regions of the network

indicates UHPLC-MS specificity, however, these features could not have been annotated with high confidence without the proper annotation strategy. The compound classes which are found in the intersection between UHPLC-MS and FT-ICR/MS are at the same time the most influential classes for the statistics.

Proper MS-data fusion was inappropriate with the given datasets. Using the sum formulas of matching features, their relevant normalized weights/loadings from the multivariate analyses could be combined by averaging. The sum formula-based matching of FT-ICR/MS and UHPLC-MS statistical datasets revealed that overall 240 FT-ICR/MS peaks were matched to 392 UHPLC-MS entries. 148 (67%) of the FT-ICR/MS features were successfully matched against HMDB 3.5, while the same was true for 257 (66%) UHPLC-MS features.

The combination of the FT-ICR/MS weights and the UHPLC-MS loadings over all HMDB matched features showed (Figure3) that glycerophospholipids dominate the IS up-regulated features. A comparison with the Figures S4 and S6 underlined the analogous behavior of the remaining compound classes. The compound classes associated to the IS down-regulated group of features encompass more polar compound classes such as amino acids, monosccharides and carboxylic acids. These findings confirm both, the results of the two separated statistics performed on FT-ICR/MS data and UHPLC-MS data, and the findings of Lehmann et al [16].

The successful RT-supported matching of accurate FT-ICR/MS data to UHPLC-MS data enabled the correct annotation of isomers that were separated by liquid chromatography. This is of great importance since the amount of possible isobars within common UHPLC-MS error ranges is large. The use of FT-ICR/MS alone would necessitate the development of targeted separation strategies and the purchase of multiple chemical standard compounds in order to carry out compound identification. A direct matching of DI-ESI-FT-ICR/MS data to the corresponding UHPLC-MS data has the following advantage: the RT-dimension of a feature that was successfully matched between both techniques and that has shown to be of statistical importance throughout these platforms, immediately discerns and locates the correct isomer to be isolated. In addition, RT information reveals the UHPLC-MS conditions

which need to be optimized in order to perform targeted analyte enrichment for either MS/MS-based compound identification or NMR-based identification after preparative analyte fraction collection.

CONCLUSION

Classical annotation workflows that do not use authentic chemical standards, treat each detected m/z peak independently from all other detected m/z species. The coverage of annotation is thereby limited by ion abundance because such combinatorial methods require validation via isotopic patterns and/or MS/MS experiments. Workflows that employ authentic chemical standards are limited by the completeness of internal standard databases. Furthermore, chemical standards can suppress the actual analytes, which results in larger required sample size. The presented approach puts all detected features into a relational network that optimizes annotations for overall consistency of all mass differences and RT-differences. Stability of annotations can be assessed by means of multiple annotation cycles, which attributes each sum formula with a specific annotation probability that is independent of arbitrary definitions of 'sufficient' mass accuracy. The presented approach is therefore a solution for high confidence feature annotation, prior to the matching of different metabolomics platforms.

Despite of the lack of corresponding retention time data it is possible to compare FT-ICR/MS data and UHPLC-MS data on the basis of *m/z* and RT information. After sum formula assignment to high accuracy FT-ICR/MS data, it is possible to discern UHPLC-MS features that (i) are specific to the analyzed sample matrix and (ii) have a high probability to share the same composition as the corresponding FT-ICR/MS feature. These characteristic compositions in the UHPLC-MS dataset can be used as starting points for mass difference network-based UHPLC-MS feature assignment. This strategy provides a solid basis for correct data matching and consecutive comparison of multivariate statistics. The presented method is of particular benefit for low accuracy MS users, yet it is a useful tool for high accuracy MS as well. Naturally, any other classical method for high confidence feature annotation can be applied for the definition of the required starting points. The presented workflow

can be performed on any column chemistry and the data driven mining of RTMDs avoids the application of false rules if column chemistry is changed. The ultimate aim of metabolomics to detect, identify and quantify all metabolites of a sample, organism or cell is more likely to be achieved by multi-platform strategies. As demonstrated in this manuscript, the intersection between the metabolomes revealed by both investigated platforms was small relative to the overall amount of detected and annotated features. On one side, MS data matching can only be performed on the intersection of two different approaches, which intrinsically limits statistical metabolome coverage. On the other side, features that were successfully matched and validated in their statistical behavior are more tractable for targeted compound identification. The costs and time to be invested for the development of quantitative targeted detection methods for disease markers might drastically decrease as a result of platform matching.

ASSOCIATED CONTENT

AUTHOR INFORMATION

Corresponding authors

sara.forcisi@helmholtz-muenchen.de

schmitt-kopplin@helmholtz-muenchen.de

Author Contributions

§These authors contributed equally.

ACKNOWLEDGMENTS

The authors would like to thank the German Federal Ministry of Education and Research (BMBF), the German Center for Diabetes Research (DZD; Grant 01GI0925), the Kompetenznetz Diabetes mellitus (Competence Network for Diabetes mellitus) funded by the German Federal Ministry of Education and Research (FKZ 01GI0804 and 01GI1104A and B).

Supporting Information Available

Additional information as noted in the text. Tables S1-S4, Figures S1-S8, supporting experimental information and supporting

concepts. This information is available free of charge via the Internet at http://pubs.acs.org/.

REFERENCES

(1) Gika, H. G.; Theodoridis, G. A.; Plumb, R. S.; Wilson, I. D. *J. Pharm. Biomed. Anal.* **2014**, *87*, 12-25.

(2) Carvalho, L. M.; Carvalho, F.; de Lourdes Bastos, M.; Baptista, P.; Moreira, N.; Monforte, A. R.; da Silva Ferreira, A. C.; de Pinho, P. G. *Talanta* **2014**, *118*, 292-303.

(3) Li, Y.; Song, X.; Zhao, X.; Zou, L.; Xu, G. *J. Chromatogr. B: Anal. Technol. Biomed. Life Sci.* **2014**, *966*, 147-153.

(4) Naz, S.; Garcia, A.; Barbas, C. *Anal. Chem.* **2013**, *85*, 10941-10948.

(5) Forcisi, S.; Moritz, F.; Kanawati, B.; Tziotis, D.; Lehmann, R.; Schmitt-Kopplin, P. *J. Chromatogr. A* **2013**, *1292*, 51-65.

(6) Smilde, A. K.; van der Werf, M. J.; Bijlsma, S.; van der Werff-van der Vat, B. J.; Jellema, R. H. *Anal. Chem.* **2005**, *77*, 6729-6736.

(7) Tikunov, Y. M.; de Vos, R. C.; Gonzalez Paramas, A. M.; Hall, R. D.; Bovy, A. G. *Plant Physiol.* **2010**, *152*, 55-70.

(8) Forshed, J.; Idborg, H.; Jacobsson, S. P. *Chemom. Intell. Lab. Syst.* **2007**, *85*, 102-109.

(9) Smolinska, A.; Blanchet, L.; Coulier, L.; Ampt, K. A.; Luider, T.; Hintzen, R. Q.; Wijmenga, S. S.; Buydens, L. M. *PloS one* **2012**, *7*, e38163.

(10) Tziotis, D.; Hertkorn, N.; Schmitt-Kopplin, P. *Eur. J. Mass Spectrom*. **2011**, *17*, 415-421.

(11) Angulo, P. *N. Engl. J. Med.* **2002**, *346*, 1221-1231.

(12) Ruhl, C. E.; Everhart, J. E. *Clin. Liver Dis.* **2004**, *8*, 501-519, vii.

(13) Browning, J. D.; Szczepaniak, L. S.; Dobbins, R.; Nuremberg, P.; Horton, J. D.; Cohen, J. C.; Grundy, S. M.; Hobbs, H. H. *Hepatology* **2004**, *40*, 1387-1395.

(14) Fassio, E.; Alvarez, E.; Dominguez, N.; Landeira, G.; Longo, C. *Hepatology* **2004**, *40*, 820-826.

(15) Stefan, N.; Haring, H. U. *Diabetes* **2011**, *60*, 2011-2017.

(16) Lehmann, R.; Franken, H.; Dammeier, S.; Rosenbaum, L.; Kantartzis, K.; Peter, A.; Zell, A.; Adam, P.; Li, J.; Xu, G.; Konigsrainer, A.; Machann, J.; Schick, F.; Hrabe de Angelis, M.; Schwab, M.; Staiger, H.; Schleicher, E.; Gastaldelli, A.; Fritsche, A.; Haring, H. U.; Stefan, N. *Diabetes care* **2013**, *36*, 2331-2338.

(17) Suhre, K.; Schmitt-Kopplin, P. *Nucleic Acids Res.* **2008**, *36*, W481-484.

(18) Wishart, D. S.; Jewison, T.; Guo, A. C.; Wilson, M.; Knox, C.; Liu, Y.; Djoumbou, Y.; Mandal, R.; Aziat, F.; Dong, E.; Bouatra, S.; Sinelnikov, I.; Arndt, D.; Xia, J.; Liu, P.; Yallou, F.; Bjorndahl, T.; Perez-Pineiro, R.; Eisner, R.; Allen, F.; Neveu, V.; Greiner, R.; Scalbert, A. *Nucleic Acids Res.* **2013**, *41*, D801-807.

(19) Vaughan, A. A.; Dunn, W. B.; Allwood, J. W.; Wedge, D. C.; Blackhall, F. H.; Whetton, A. D.; Dive, C.; Goodacre, R. *Anal. Chem.***2012**, *84*, 9848-9857.

(20) Morreel, K.; Saeys, Y.; Dima, O.; Lu, F.; Van de Peer, Y.; Vanholme, R.; Ralph, J.; Vanholme, B.; Boerjan, W. *Plant Cell* **2014**, *26*, 929-945.

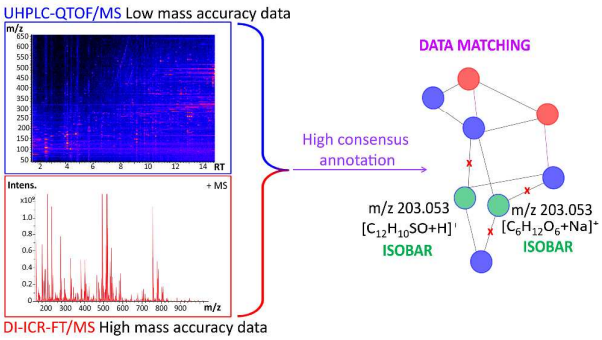
(21) Wiklund, S.; Johansson, E.; Sjostrom, L.; Mellerowicz, E. J.; Edlund, U.; Shockcor, J. P.; Gottfries, J.; Moritz, T.; Trygg, J. *Anal. Chem.* **2008**, *80*, 115-122.

(22) Subramanian, A.; Tamayo, P.; Mootha, V. K.; Mukherjee, S.; Ebert, B. L.; Gillette, M. A.; Paulovich, A.; Pomeroy, S. L.; Golub, T. R.; Lander, E. S.; Mesirov, *J. P. Proc. Natl. Acad. Sci. U. S. A.* **2005**, *102*, 15545-15550.

(23) Jansson, J.; Willing, B.; Lucio, M.; Fekete, A.; Dicksved, J.; Halfvarson, J.; Tysk, C.; Schmitt-Kopplin, P. *PloS one* **2009**, *4*, e6386.

(24) Lucio, M.; Fekete, A.; Weigert, C.; Wagele, B.; Zhao, X.; Chen, J.; Fritsche, A.; Haring, H. U.; Schleicher, E. D.; Xu, G.; Schmitt-Kopplin, P.; Lehmann, R. *PloS one* **2010**, *5*, e13317.

For TOC only


UHPLC-QTOF/MS Low mass accuracy data
m/z
RT
Intens.
+ MS
m/z
DI-ICR-FT/MS High mass accuracy data
High consensus annotation
DATA MATCHING
m/z 203.053
[C12H10SO+H]+
ISOBAR
m/z 203.053
[C6H12O6+Na]+
ISOBAR

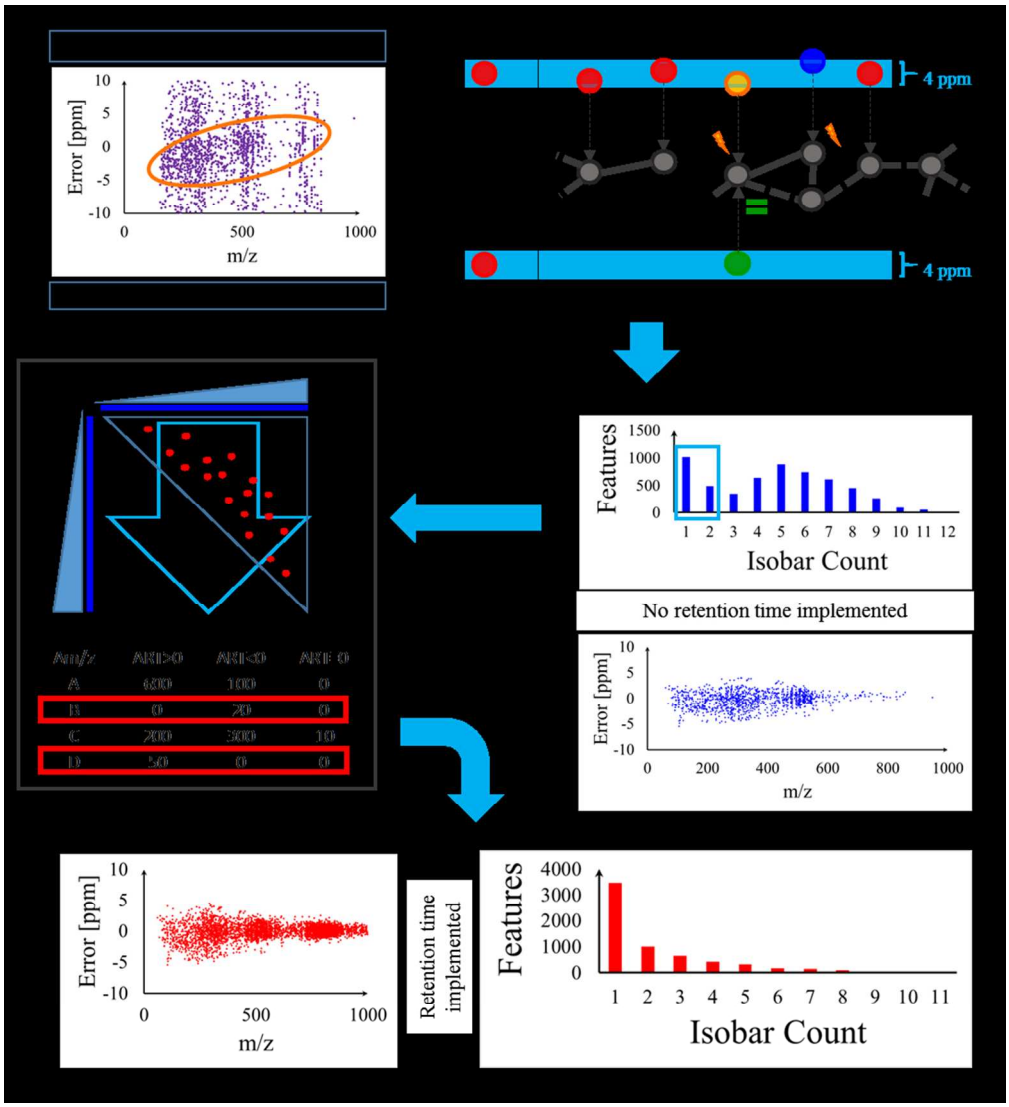


Figure 1. Convolution of the RT-supported UHPLC-TOF-MS annotation strategy. A) Error over m/z plot of theoretical FT-ICR/MS ion masses to UHPLC-TOF-MS data matching and reasoning behind the selection of biunique Hits for MDiN annotation. The isomeric annotation of possible isobars may project multiple false starting points into the MDiN. B) Summary of 100 repetitive MDiN annotations. More than 80% of all features obtain multiple isobaric annotations. The error over m/z distribution of the stable annotations is well centered. Features with m/z > 500 are poorly annotated. C) The stable annotations, their RTs and Formulas are screened for RT-specific mass differences. Only mass differences with absolute RT+/-- specificity (at least 20 vs. zero observations) are considered for MDiN reconstruction. D) Summary of 100 repetitive RT-MDiN annotations. The number of stable annotations is increased more than 3-fold; the proportion of unstable annotation is decreased. The corresponding error over m/z distribution is well centered and features with m/z > 500 are well addressed.
183x201mm (150 x 150 DPI)

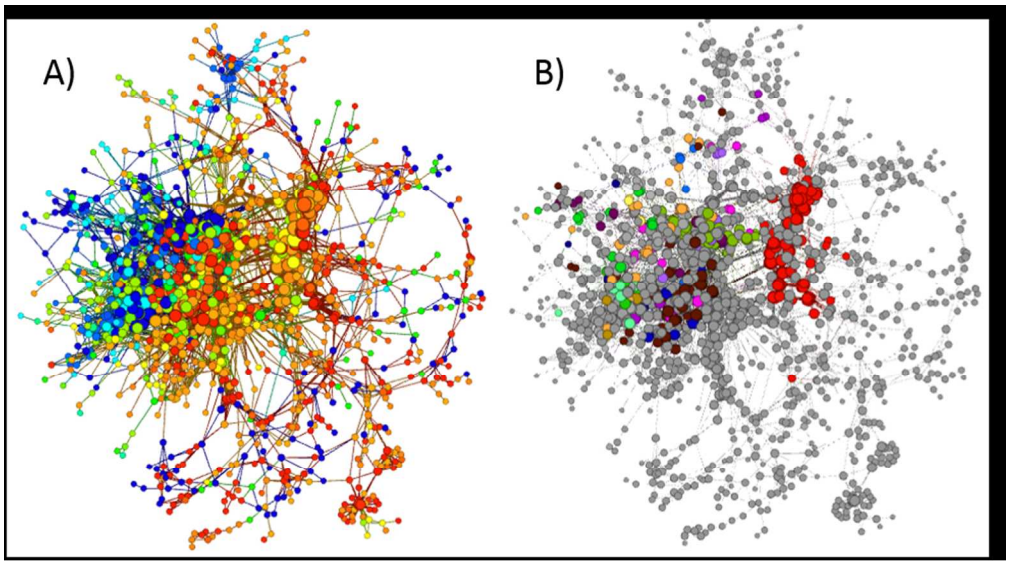


Figure 2. MDiN of annotated UHPLC-TOF-MS features in three different layouts: A) RT-gradient from blue over green to red. B) Glycerophospholipids (red), Prenol Lipids (brown), Carboxylic Acids (dark purple), Glycosyl Compounds (green), Amino Acids (orange), Fatty Alcohols (purple), Fatty Acids and Conjugates (pink), Fatty Acid Esters (olive), Fatty Amides (violet), Steroids and Derivateves (blue), Monosaccharides (yellow), Eicosanoids (light blue). Compound Classes with lower frequency or missing HMDB annotation are grey.
166x92mm (150 x 150 DPI)

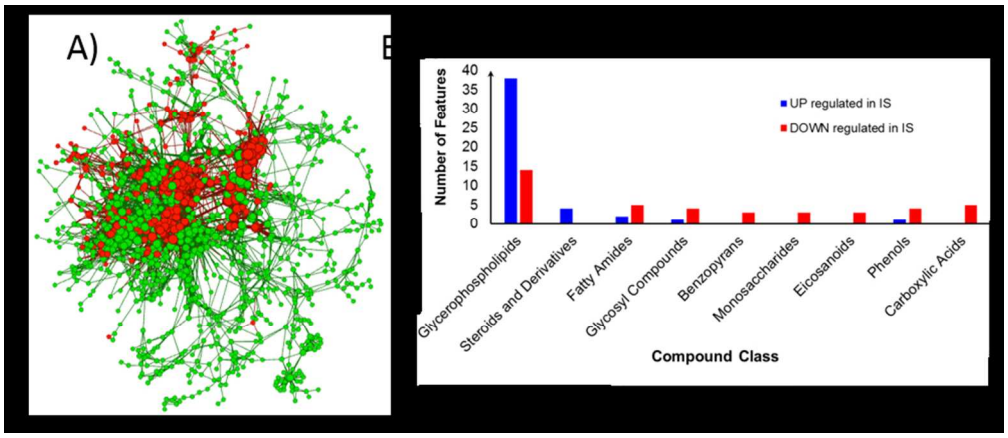


Figure 3. A) Positioning of corresponding FT-ICR/MS-UHPLC-TOF-MS features (red). B) Compound class distribution of the combined statistics. Up-regulated in IS is blue and down-regulated in IS is red.
168x71mm (150 x 150 DPI)