eRah: A Computational Tool Integrating Spectral Deconvolution and Alignment with Quantification and Identification of Metabolites in GC–MS-Based Metabolomics

Xavier Domingo-Almenara,[*,†,‡] Jesus Brezmes,[†,‡] Maria Vinaixa,[†,‡] Sara Samino,[†,‡] Noelia Ramirez,[†,‡] Marta Ramon-Krauel,[¶] Carles Lerin,[¶] Marta Díaz,[¶,‡] Lourdes Ibáñez,[¶,‡] Xavier Correig,[†,‡] Alexandre Perera-Lluna,[§] and Oscar Yanes[*,†,‡]

†*Metabolomics Platform, Department of Electronic Engineering (DEEEA), Universitat Rovira i Virgili, Tarragona, Catalonia, Spain*

‡*Biomedical Research Centre in Diabetes and Associated Metabolic Disorders (CIBERDEM), Madrid, Spain*

¶*Institut de Recerca Pediàtrica, Hospital Sant Joan de Déu, University of Barcelona, Barcelona, Catalonia, Spain*

§*B2SLab, Center for Biomedical Engineering Research (CREB), CIBERBBN, Department of ESAII, Universitat Politècnica de Catalunya, Barcelona, Catalonia, Spain*

E-mail: xavier.domingo@urv.cat; oscar.yanes@urv.cat

Abstract

Gas chromatography coupled to mass spectrometry (GC–MS) has been a long-standing approach used for identifying small molecules due to the highly reproducible ionization process of electron impact ionization (EI). However, the use of GC–EI MS in untargeted metabolomics produces large and complex datasets characterized by co-eluting compounds and extensive fragmentation of molecular ions caused by the hard electron ionization. In order to identify and extract quantitative information of metabolites across multiple biological samples, integrated computational workflows for data processing are needed. Here we introduce eRah, a free computational tool written in the open language R composed of five core functions: (i) noise filtering and baseline removal of GC–MS chromatograms, (ii) an innovative compound deconvolution process using multivariate analysis techniques based on compound match by local covariance (CMLC) and orthogonal signal deconvolution (OSD), (iii) alignment of mass spectra across samples, (iv) missing compound recovery, and (v) identification of metabolites by spectral library matching using publicly available mass spectra. eRah outputs a table with compound names, matching scores and the integrated area of compounds for each sample. The automated capabilities of eRah are demonstrated by the analysis of GC-TOF MS data from plasma samples of adolescents with hyperinsulinaemic androgen excess and healthy controls. The quantitative results of eRah are compared to centWave, the peak-picking algorithm implemented in the widely used XCMS package, MetAlign and ChromaTOF software. Significantly dysregulated metabolites are further validated using pure standards and targeted analysis by GC-QqQ MS, LC-QqQ and NMR. eRah is freely available at http://CRAN.R-project.org/package=erah.

Introduction

Metabolomics is widely used to obtain new insights into human, plant and microbial biochemistry, as well as in drug discovery, nutrition research and food control. Although different technologies are nowadays used to achieve these objectives,[1] the proof of concept for what we now know as mass spectrometry-based metabolomics was reported in 1966 by Dalgliesh et al.,[2] which conducted the first GC–MS experiment to separate a wide range of metabolites occurring in urine and tissue extracts. Later in 1971, Horning et al.[3] introduced the term "metabolic profiles", and along with Pauling and Robinson led to the development of GC–MS methods for monitoring metabolites in biological samples through the 1970s.[4,5]

GC–MS has been a long-standing approach used for metabolite profiling of volatile and semi-volatile compounds due to the widespread use of electron impact ionization (EI). EI is a hard ionization technique that has been historically standardized at 70 eV. Unlike soft ionization techniques such as ESI[6] or MALDI,[7] EI is a highly reproducible ionization process across many different platforms. However, co-elution of compounds from complex biological samples in GC along with extensive fragmentation of molecular ions by EI ionization, result in large and complex datasets. Reconstructing GC–MS profile data into identified and quantified metabolites across multiple samples remains a challenging task due to the lack of integrated computational tools in GC–MS-based untargeted metabolomics.

Current computational approaches for GC–MS data processing fall into two main categories: tools based on peak-picking, and tools for compound extraction through the so-called curve resolution or spectral deconvolution. The first category involves detecting all relevant fragment ion peaks in the spectra, and align them across multiple samples[8,9] to subsequently discover statistical peak variations between experimental groups. Representative tools from this category include MZmine,[10,11] MetAlign,[12,13] and XCMS.[14,15] Although these tools were initially intended for liquid chromatography mass spectrometry (LC–MS) data processing, they can also be used for GC–MS data analysis.[16,17] The quantitative variables provided by these methods are not based on the compound spectra, but the m/z value, retention time

window and area of fragment ion peaks. Thus, compound identification is the main bottleneck of peak-picking approaches. In this regard, tools such as metaMS,[18] TagFinder,[19] MetaboliteDetector[20] and PyMS[21] attempt to overcome this limitation by grouping the different peaks (based on their shape similarity or peak correlations) into partial compound spectra, allowing the putative identification of compounds by comparing their mass spectra with a reference MS library.

The second category focuses on the compound as the analysis entity, as opposed to the use of individual fragment ion peaks. Compounds are quantified and identified on the basis of a multivariate deconvolution process[22] that extracts and constructs pure compound spectra from raw data. Representative tools falling into this category include TNO-DECO[23] or ADAP-GC.[24] TNO-DECO uses multivariate curve resolution to extract the compound spectra, whereas the deconvolution algorithm of ADAP-GC is based on an hierarchical clustering of fragment ions. Other free software, such as AMDIS[25] or BinBase[26,27] perform parts of the GC–MS metabolomics workflow. AMDIS is used to identify compounds by using the NIST library, but it processes samples independently and it does not include spectral alignment. BinBase uses the spectral deconvolution provided by a proprietary algorithm in the commercial software ChromaTOF (LECO Corporation) in order to align compounds across samples, and it provides compound quantification and identification based on self-constructed libraries.[28]

Despite these efforts, there is a need for a free and open source software that integrates all the necessary steps for data processing in GC–MS-based untargeted metabolomics. Here we introduce eRah, an R package with an integrated design that incorporates a novel spectral deconvolution method using multivariate techniques based on blind source separation (BSS), alignment of spectra across samples, quantification, and automated identification of metabolites by spectral library matching. We demonstrate the functionality of eRah through a comparative analysis of serum samples from adolescents with hyperinsulinaemic androgen excess (HIAE) and healthy controls.

Experimental Section

Materials

A dataset of 25 serum samples (from 11 young, non-obese adolescents with HIAE and 14 age-, weight- and ethnicity-matched healthy controls)[29] were analyzed by GC–EI-qTOF-MS (Agilent Technologies). A second cohort of 74 plasma samples from healthy individuals were analyzed by GC–EI-TOF-MS (Pegasus, LECO Corporation). Pure standards nicotinic acid, leucine, proline, methionine, aspartic acid, myo-inositol, ornithine, urea and lactic acid were purchased from Sigma Aldrich (Steinheim, Germany). Analytical grade methanol was purchased from SDS (Peypin, France). Water was produced in an in-house Milli-Q purification system (Millipore, Molsheim, France). N-methyl-N-trimethylsilyltrifluoroacetamide, methoxamine hydrochloride and pyridine were purchased from Sigma-Aldrich (Steinheim, Germany). Myristic-d27 acid and succinic acid-2,2,3,3-d4 were from Isotec Stable Isotopes (Miamisburg, USA).

Metabolite extraction method

Serum aliquots (25 μL) were thawed at 4 °C. Samples were briefly vortex-mixed and each aliquot was supplemented with 20 μL of 1 μg/μL succinic-d4 acid (internal standard). Proteins were then precipitated by the addition of 475 μL cold methanol/water (8:1 vol/vol) followed by 3 min of ultrasonication and 10 s of vortex-mixing. Aliquots were subsequently maintained on ice for 10 min. After centrifugation for 10 min (19.000 g, 4 °C), 100 μL of supernatant were transferred to a GC autosampler vial and lyophilized. We incubated the lyophilized serum residues with 50 μL methoxyamine in pyridine (40 μg/μL) for 30 min at 60 °C. To increase the volatility of the compounds, we silylated the samples using 30 μL N-methyl-N-trimethylsilyltrifluoroacetamide with 1% trimethylchlorosilane (Thermo Fisher Scientific) for 30 min at 60 °C.

GC–TOF MS analysis

Analysis was carried out on a qTOF MS 7200 (Agilent, Santa Clara, CA, USA) coupled to an Agilent 7890A gas chromatography (GC). Derivatized samples (1 μL each) were injected in the gas chromatograph system with a split inlet equipped with a J&W Scientific DB5–MS+DG stationary phase column (30 mm $\times$ 0.25 mm i.d., 0.1 μm film, Agilent Technologies). Helium was used as a carrier gas at a flow rate of 1 mL/min in constant flow mode. The injector split ratio was adjusted to 1:5 and oven temperature was programmed at 70 °C for 1 min and increased at 10 °C/min to 325 °C. The MS was operated in the electron impact ionization mode at 70 eV. Mass spectral data were acquired in full scan mode from m/z 35 to 700 with an acquisition rate of 5 spectra per second. Details on the Pegasus GC-TOF (LECO Corporation) and GC-QqQ MS analyses are available in the Supporting Material.

Data processing methods

With the aim of comparing the quantitative results of GC–MS serum samples, the data set was processed using eRah, XCMS,[14,15] MetAlign[13] and ChromaTOF (LECO Corp.) software. GC–MS data files were converted to .mzXML format using Proteowizard software.[30] Converted files were processed using XCMS and MetAlign in order to detect and align features. A feature is defined as an ion entity with a unique m/z and a specific retention time (mzRT). The parameters used in the XCMS workflow were: `xcmsSet (method = 'centWave', ppm = 15, peakwidth = c(1,5)); retcor (method = 'peakgroups', extra = 1, missing=1)` and `group (mzwid = 0.0025, minfrac = 0.5, bw = 5)`. XCMS analysis provided an xcmsSet object containing the retention time, m/z value, and peak intensity (or area) of each feature for every serum sample. MetAlign parameters were as default, with an average peak width of 1 s. ChromaTOF software settings included a baseline offset of 1, mass threshold of 10, and peaks were required to have a minimum similarity score of 600 before assigning a name. Unique mass was used for peak area calculation and LECO's statistical compare module was used

to align compounds across samples. Converted files were also processed using eRah through a fast script in R, which includes (i) data pre-processing, (ii) spectral deconvolution, (iii) spectral alignment, (iv) missing compound recovery, and (v) compound identification (see details below). The samples raw-data are classified in folders, where each folder is a class. Signals at m/z 73, 74, 75, 147, 148, and 149 were excluded for data processing, since these are ubiquitous mass fragments typically generated from compounds carrying a trimethylsilyl moiety. We used the mass range 70-600 m/z (except for the six excluded m/z) for comparison between deconvoluted and reference spectra. Note that selected/excluded masses can be modified according to the user criterion. eRah does not use excluded m/z for deconvolution, alignment and identification, and it sets those library's masses to zero, so it does not affect spectral matching and identification. If, instead, the user decides not to exclude any m/z, eRah compares the full range of masses in the identification step. The eRah parameters for the deconvolution were: `setDecPar(min.peak.width=1, min.peak.height=2000, noise.threshold=500, avoid.processing.mz= c(73:75,147:149))`, and for the alignment: `setAlPar(min.spectra.cor=0.90, max.time.dist=3, mz.range=1:600)`. The minimum number of samples was set to 8 for the missing compound recovery step. The complete analysis of 25 and 74 samples was performed in less than 30 and 90 minutes respectively (in a 2.4 GHz Intel Core i7 computer). The eRah package includes a tutorial and the description of each function and parameter through the R help.

Results and discussion

Computational workflow

This section describes the five steps of the eRah workflow (Figure 1): (i) data pre-processing, (ii) spectral deconvolution, (iii) spectral alignment, (iv) missing compound recovery, and (v) compound identification. A detailed explanation of eRah methods can be found in the

Supporting Information.

(i) Pre-processing.

GC–MS chromatograms are usually affected by baseline drift and instrumental noise. Smoothing the data by noise filtering and baseline removal improves the eRah's deconvolution and alignment algorithms. Both baseline and noise are filtered according to a minimum compound peak width σ_{MIN}, a value (in seconds) selected by the user. eRah then approximates the baseline drift by a moving-minimum filter[31] to correct the chromatogram, and removes noise using Savitzky-Golay filter.[32]

(ii) Deconvolution.

eRah performs a two-step compound deconvolution. First, a multivariate matched filter called compound match by local covariance (CMLC) is applied. The CMLC filter is based on the covariance match filter[33] applied using local covariance matrices.[34] This multivariate approach benefits from the inherent correlation of fragment ions of each compound in EI-MS. CMLC uses covariance matrices to detect patterns of ion redundancy that characterize each compound within the chromatogram. The patterns of ion redundancy approximate to a gaussian peak shape. This matched filter outputs a signal with local minima on spots of ion redundancy in the chromatogram, which are determined by compounds with a peak width equal or greater than the selected σ_{MIN} (Figure 2(a)). Upon compound detection by CMLC, the pure compound spectrum is determined using a blind source separation-based algorithm known as orthogonal signal deconvolution (OSD).[35,36] OSD is a method able to retrieve a compound spectrum given a compound elution profile. We approximate the elution profile for OSD with the same gaussian model used in CMLC. After the spectrum is determined, we obtain the quantitative compound profile with a least absolute deviation (LAD) regression[37] between the spectrum found by OSD and the chromatogram. It is worth highlighting that this tandem compound match–OSD strategy in eRah performs a multivariate spec-

tral deconvolution that avoids the estimation of the number of factors/components, a key parameter that affects the outcome of traditional multivariate algorithms used for spectral deconvolution, including multivariate curve resolution or parallel factor analysis.

This two-step deconvolution in eRah is depicted in Figure 2 using a mixture of five standard compounds, where two different co-elution scenarios are shown. Figures 2(b) and 2(c) show the eRah's resolved chromatograms for nicotinic acid (I), isoleucine (II) and proline (III) (minutes 5.65–5.74), and methionine (IV) and aspartic acid (V) (minutes 7.13–7.19), respectively. The five compounds were detected using CMLC and their corresponding spectra successfully deconvolved by OSD. In case of processing accurate m/z data (e.g., Agilent GC-qTOF MS), compounds are deconvoluted to nominal mass.

(iii) Alignment.

This step aims to correct the retention time variation of the eluting compounds, facilitating the relative quantification and comparison of compounds across samples. Firstly, the user selects the maximum retention time drift (in seconds) and the minimum spectral similarity (from 0 to 1, being 0 no similarity and 1 the highest similarity) that will be allowed for alignment. This means that two or more compounds with a retention time distance above a maximum retention time drift are not aligned because they are seen as different compounds. The same occurs with the minimum spectral similarity. The alignment is then performed by clustering compounds within these boundaries of retention time distance and spectral similarity (Figure 3(a) and Supporting Information). To determine these clusters eRah computes the Euclidean distance between retention time distance and spectral similarity for all compounds in the chromatograms, resulting in compounds appearing across the maximum number of samples and with the least retention time and spectral distance. As an indicative example, Figure 3(b) shows the profile of urea before and after the alignment step. Our method, therefore, does not require any internal reference, such as retention indices, to align compounds across samples. Nevertheless the user still has the possibility to use retention

indices to increase confidence in compound identification. The alignment algorithm has been implemented for a high run-time performance and it is capable of aligning large datasets. In this regard, we have processed successfully up to 1.200 samples (data not shown).

(iv) Missing Compounds Recovery.

Alignment in eRah is a blind step where the algorithm is not forced to find compounds throughout the samples. This means that, in certain circumstances where a strong variation in a compound spectrum occurs, for instance, due to low concentration in a sample (leading noise to disturb the compound spectrum), the alignment step may fail to group the same compound in all samples. To resolve this situation we have implemented a missing compound recovery step. Similarly to XCMS, the user may impose that compounds appearing in at least e.g., 80% of the samples in an experimental class, may also be found in all other samples. To do this, eRah determines a target spectrum from the mean spectrum of each aligned compound. As in the deconvolution step, eRah retrieves the compound chromatographic profile by a LAD regression between the target spectrum and a chromatographic window around the expected elution time in the samples where the compound is missing.

(v) Identification.

Aligned compounds are identified by comparing the mean spectrum to reference spectra in a MS library.[38] The mean spectrum is determined by the mean of the compounds spectra found only in the deconvolution and alignment steps. The current eRah package integrates the free and downloadable version of the MassBank[39] repository, which after removing duplicated compounds contains a set of ∼500 unique compounds with EI GC–TOF mass spectra. However, users may import other libraries such as the Golm Metabolome Database (GMD),[40,41] Fiehn,[28] Human Metabolome Database (HMDB)[42] or an internal database, as long as the library is available in an interpretable (i.e., readable) format. This refers to a non-binary non-coded format, which is consistent with public databases. By comparing the empirical

spectra with a reference MS library, eRah generates a list of candidate metabolites along with a similarity match factor, determined using the cosine product[43] (see Supporting Information for details). Note that NIST library can only be read with NIST software. However, eRah has a function to export all the spectra found in a given experiment to MSP format. Spectra in MSP format can be imported to the NIST MS Search software, allowing the comparison of the eRah's spectra with the NIST library.

Comparative analysis of serum samples from adolescents with hyperinsulinaemic androgen excess and healthy controls

To illustrate the integrative workflow of eRah, we carried out a comparative metabolomic analysis using 11 serum samples from girls with hyperinsulinemic androgen excess (HIAE), and 14 age-, weight- and ethnicity-matched healthy controls.[29] HIAE in post-menarcheal adolescent girls is recognized as the phenotypic core of a broader pathological entity traditionally known as polycystic ovary syndrome (PCOS),[44] which affects 8–21% of women of reproductive age worldwide.[45,46] HIAE usually precedes a broader pathological phenotype in adulthood that is associated with anovulatory infertility, metabolic syndrome, type 2 diabetes[47] and possibly cardiovascular disease.[48] Therefore, unveiling metabolic derangements in early stages can bring a better understanding of these long-term health risks.

Samples were analyzed using an Agilent GC–EI-qTOF-MS (see Methods section for further details). Raw GC–MS files are available at MetaboLights with accession number MTBLS321. With the aim of comparing the quantitative results of the deconvolved compounds by eRah, mass spectra were also processed using XCMS (Supporting File 1 and 2) and MetAlign (Supporting File 3). XCMS uses centWave, a highly sensitive peak detection algorithm.[49] The output of eRah contained 169 resolved and aligned compounds (Supporting File 4). We focused, however, on 33 compounds to assess the relative quantitative accuracy of eRah by comparison with XCMS and MetAlign (Table 1). These compounds showed a high similarity match factor (>80.0) to reference MS spectra in the GMD and MassBank.

We manually selected a quantitative mzRT feature from the XCMS and MetAlign output for each of the 33 compounds. Given the multivariate nature of the spectral deconvolution in eRah, compound relative quantification is based on the area of the deconvolved compound elution profile and not just a fragment ion peak. Table 1 shows the list of 33 compounds with their retention time (RT), identification match factor (MF), and quantitative ion from XCMS and MetAlign. To demonstrate that eRah performs well in a wide dynamic range of metabolite concentrations, we determined the relative compound concentrations (Rel. C.) defined as the quotient between the mean concentration of each compound (i.e., mean area of each deconvolved compound profile) and the mean concentration of all the compounds listed in the table. In addition, the table shows the coefficient of determination (R^2) of the regression between the mean area - and intensity - of the deconvolved compound profile (eRah) and the quantitative mzRT feature (XCMS and MetAlign). Finally, the percentage of variation between disease (HIAE) and control was also calculated for each compound.

The analysis indicated an excellent linear correlation (R^2>0.90) for most compounds. Even for coeluted (e.g., glycerol and phosphoric acid) and low relative concentration compounds (e.g., myo-inositol and uric acid), the correlation between the area - and intensity - of deconvolved compounds and selective mzRT features was high. Only the area of hydroxylamine and lysine showed R^2<0.80, however these two compounds exhibited similar percentages of variation between HIAE and control groups when compared to XCMS. We also noted that for some compounds the coefficients of determination varied when comparing area and intensity. We attribute these differences to the fact that eRah, XCMS and MetAlign use distinct methodology for quantifying compounds and peaks, respectively, which may lead to some disagreements when comparing areas linearly. Moreover, although XCMS and MetAlign are two very reliable references, their results should not be taken as ground truth. For this reason, we validated eRah's results using two additional analytical platforms (Table 2). Due to availability of pure standards in our laboratory, we focused on lactate, myo-inositol, urea and ornithine for validation experiments. Manual integration of peaks using

MassHunter (Agilent Technologies) revealed similar relative differences, and targeted analysis using GC-triple quadrupole (QqQ) MS (see Supporting Material for details) reproduced similar variations between HIAE and control. Altogether, Table 2 consistently shows similar relative quantitative differences using eRah, XCMS, MetAlign, MassHunter and QqQ MS analysis, which further supports the strength of eRah for GC–MS-based untargeted metabolomics studies. To demonstrate further results on a larger number of samples, we processed 74 additional plasma samples analyzed using a LECO GC-TOF MS (see Supporting Material for details), and we compared the quantification of 25 metabolites with MetAlign and ChromaTOF software (see Supporting Table S-1). The supporting table shows the identification match factor and the coefficient of determination (R^2) of the regression between the mean area of the deconvolved compound profile (eRah) and a selective m/z automatically quantified by ChromaTOF. Again, the analysis indicated an excellent linear correlation (R^2>0.90) for most compounds and with both software. Only some compounds showed good correlation with either MetAlign (e.g., urea, alanine, uric acid) or ChromaTOF (e.g., valine, serine). Only the area of hydroxylamine, glycine and threonine showed R^2<0.80.

Finally, we determined statistical significant differences between HIAE and control using eRah. Lactic acid, leucine, serine, 5-oxoproline (pyroglutamic acid), glutamic acid, ornithine and lysine showed higher levels (p-value<0.01) in HIAE relative to control serum samples (Table 3). Next, we focused on changes in 5-oxoproline, glutamic acid, lactic acid and leucine to be validated by complementary analyses on the same serum samples using nuclear magnetic resonance (NMR) or liquid chromatography (LC-QqQ MS) as previously described[29] (Figure 4). Metabolites in NMR spectra were quantified using Dolphin.[50,51] Interestingly, analytical platforms such as NMR, which analyze serum non-destructively, and LC–MS, which produces intact molecular ions due to soft ionization, revealed very similar percentages of variation and p-values. Moreover, Zhao et al.[52] observed a positive association of lactate and leucine concentrations with insulin resistance independently of obesity in adult (28-29 years old) PCOS patients. In this previous study, serine levels were increased in PCOS

plasma samples as compared with the normal controls independently of obesity or insulin resistance.[52] Our results also revealed elevated levels of lactate, leucine and serine, although in non-obese adolescents with hyperinsulinaemic androgen excess. Finally, elevated level of ornithine suggests the imbalance of urea cycle in adolescents with HIAE.

Altogether, our results strengthen the feasibility of the recently challenged[53] GC–MS technique for metabolomic applications, and demonstrate the robustness of eRah for data processing.

Conclusions

Despite the existence of different pieces of free and commercial software for GC–MS data analysis, none of these allow the execution of an integrated workflow that includes spectral deconvolution and alignment, followed by the identification and quantification of metabolites in the same application. This still leads many researchers to implement separate software for each process, and tedious manual workflows for data processing. We have developed eRah to fill this gap. eRah is a free computational tool written in the open language R and deposited at a public and permanent repository (http://CRAN.R-project.org/package=erah), that enables users to execute a complete automated workflow for data analysis in GC–MS untargeted metabolomics. Moreover, eRah incorporates an innovative deconvolution process based on multivariate compound detection and blind source separation that differs from existing tools. Besides, the innovative modular and open-structure of eRah allows for the implementation of new algorithms by the scientific community, facilitating reproducibility and comparison between algorithms and methods. The comparative analysis of serum samples by GC–TOF MS provided test data demonstrating an excellent correlation with alternative quantitative approaches and analytical platforms such as XCMS, MetAlign, ChromaTOF, GC–QqQ MS, LC–QqQ MS and NMR, with the manifest advantage that eRah provides a complete automated data analysis workflow. Collectively, we anticipate that eRah will help

to expedite and facilitate the analysis of GC–MS data resulting in a greater implementation of this technique in untargeted metabolomic studies.

Acknowledgement

The authors acknowledge Dr. R. Ras and Mrs. S. Mariné from the Centre for Omic Sciences for providing technical assistance, and Mr. J. Gómez from the Institut d'Investigació Sanitària Pere Virgili (IISPV) for helping in the quantification of metabolites in NMR spectra. This research was partially funded by MINECO grant TEC2012-31074 and TEC2015-69076-P (to XC), TEC2013-44666-R and TEC2014-60337-R (to AP) and SAF2011-30578 and BFU2014-57466 (to OY). CIBER-BBN and CIBER-DEM are initiatives of the Spanish Instituto de Salud Carlos III (ISCIII). NR acknowledges the financial support of the European Union's Horizon 2020 research and innovation program under the Marie Sklodowska-Curie grant agreement (No. 660034). XD also acknowledges the Martí i Franquès grant from URV (2012BPURV-43) for the financial support.

Supporting Information Available

Supporting methods are included. From the processing of the HIAE experiment, XCMS detected the m/z features and quantified them by area (Supporting File 1) and intensity (Supporting File 2), and the MetAlign detected the m/z features and quantified them by area (Supporting File 3). eRah instead, provided a list of putative identifications with their respective relative concentration (Supporting File 4).

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

References

(1) Patti, G. J.; Yanes, O.; Siuzdak, G. *Nat. Rev. Mol. Cell Biol.* **2012**, 13, 63–269.

(2) Dalgliesh, C. E.; Horning, E. C.; Horning, M. G.; Knox, K. L.; Yarger K. *Biochem. J.* **1966**, 101, 792–810.

(3) Horning, E. C.; Horning, M. G. *Clin. Chem.* **1971**, 17, 802–809.

(4) Teranishi, R.; Mon, T. R.; Robinson, A. B.; Cary, P.; Pauling, L. *Anal. Chem.* **1972**, 44, 18–20.

(5) Matsumoto, K. E.; Partridge, D. H.; Robinson, A. B.; Pauling, L.; Flath, R, A.; Mon, T. R.; Teranishi, R. *J. Chromatogr. A.* **1973**, 85, 31–34.

(6) Fenn, J. B. *Angew. Chem. Int. Ed. Engl.* **2003**, 42, 3871–3894.

(7) Karas, M.; Ralf, K. *Chem. Rev.* **2003**, 103, 427–440.

(8) Koh, Y.; Pasikanti K. K.; Yap, C. W.; Chan, E. C. *J. Chromatogr. A.* **2010**, 1217, 8308–8316.

(9) Niu, W.; Knight, E.; Xia, Q.; McGarvey, B.D. *J. Chromatogr. A.* **2014**, 1374, 199–206.

(10) Katajamaa, M.; Jarkko, M.; Matej, O. *Bioinformatics.* **2006**, 22, 634–636.

(11) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Oresic, M. *BMC Bioinformatics.* **2010**, 11, 395.

(12) Lommen, A. *Anal. Chem.* **2009**, 81, 3079–3086.

(13) Lommen, A.; Harrie J. K. *Metabolomics.* **2012**, 8, 719–726.

(14) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. *Anal. Chem.* **2006**, 78, 779–787.

(15) Tautenhahn, R.; Patti, G. J.; Rinehart, D.; Siuzdak, G. *Anal. Chem.* **2012**, 84, 5035–5039.

(16) Aggio, R.; Villas-Boas, S. G.; Ruggiero, K. *Bioinformatics.* **2011**, 27, 2316–2318.

(17) Fernandez-Varela, R.; Tomasi, G.; Christensen J. H. *J. Chromatogr. A.* **2015**, 1384, 133–141.

(18) Wehrens, R.; Georg, W.; Fulvio, M. *J. Chromatogr. B.* **2014**, 966, 109–116.

(19) Luedemann, A.; Strassburg, K.; Erban, A.; Kopka, J. *Bioinformatics.* **2008**, 24, 732–737.

(20) Hiller, K.; Hangebrauk, J.; Jager, C.; Spura, J.; Schreiber, K.; Schomburg, D. *Anal. Chem.* **2009**, 81, 3429–3439.

(21) O'Callaghan, S.; De Souza, D. P.; Isaac, A.; Wang, Q.; Hodkinson, L.; Olshansky, M.; Erwin, T.; Appelbe, B.; Tull, D. L.; Roessner, U.; Bacic, A.; McConville, M. J.; Likic, V. A. *BMC bioinformatics.* **2012**, 13, 115.

(22) Du, X.; Steven H. Z. *Comput. Struct. Biotechnol. J.* **2013**, 4, 1–10.

(23) Jellema, R. H.; Krishnan, S.; Hendriks, M. M. W. B.; Muilwijk, B.; Vogels J. T. W. E. *Chemometr. Intell. Lab.* **2010**, 104, 132–139.

(24) Ni, Y.; Qiu, Y.; Jiang, W.; Suttlemyre, K.; Su, M.; Zhang, W.; Jia, W.; Du, X. *Anal. Chem.* **2012**, 84, 6619–6629.

(25) Stein, S. *J. Am. Soc. Mass Spectrom.* **1999**, 10, 770–781.

(26) Fiehn,O., Wohlgemuth,G., Scholz,M. *Proc. Lect. Notes Bioinformatics.* **2005**, 3615, 224–239.

(27) Skogerson, K.; Wohlgemuth, G.; Barupal, D. K.; Fiehn, O. *BMC Bioinformatics.* **2011**, 12, 321.

(28) Kind, T.; Wohlgemuth, G.; Lee do, Y.; Lu, Y.; Palazoglu, M.; Shahbaz, S.; Fiehn, O. *Anal. Chem.* **2009**, 81, 10038–10048.

(29) Samino, S.; Vinaixa, M.; Díaz, M.; Beltran, A.; Rodríguez, M. A.; Mallol, R.; Heras, M.; Cabre, A.; Garcia, L.; Canela, N.; Zegher, F.; Correig, X.; Ibàñez, L.; Yanes, O. *Sci. Rep.* **2015**, 5, 11496.

(30) Kessner, D.; Chambers, M.; Burke, R.; Agus, D.; Mallick, P. *Bioinformatics.* **2008**, 24, 2534–2536.

(31) Pitas, J. *IEEE Trans. Circuits Syst.* **1989**, 36, 795–804.

(32) Savitzky, A.; Golay, M. J. E. *Anal. Chem.* **1964**, 36, 1627–1639.

(33) Chang C. I. *Springer Science+Bussiness Media New York.* **2003**.

(34) Caefer C. E.; Rotman S. R. *First Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing.* **2009**, 1–4.

(35) Domingo-Almenara, X.; Perera, A.; Ramirez, N.; Canellas, N.; Correig, X.; Brezmes, J. *J. Chromatogr. A.* **2015**, 1409, 226–33.

(36) Domingo-Almenara, X.; Perera, A.; Ramirez, N.; Brezmes, J. *Comput. Methods Programs Biomed.* **2016**, 130, 135–141.

(37) Li, Y.; Gonzalo, A. R. *EURASIP J. Adv. Signal Process.* **2004**, 12, 1762-1769.

(38) Vinaixa, M.; Schymanski, E .L.; Neumann, S.; Navarro, M.; Salek, R. M.; Yanes, O. *Trends Analyt. Chem.* **2016**, 78, 23–35.

(39) Horai, H.; Arita, M.; Kanaya, S.; Nihei, Y.; Ikeda, T.; Suwa, K.; Ojima, Y.; Tanaka, K.; Tanaka, S.; Aoshima, K.; Oda, Y.; Kakazu, Y.; Kusano, M.; Tohge, T.; Matsuda, F.; Sawada, Y.; Hirai, MY.; Nakanishi, H.; Ikeda, K.; Akimoto, N.; Maoka, T.; Takahashi, H.; Ara, T.; Sakurai, N.; Suzuki, H.; Shibata, D.; Neumann, S.; Iida, T.; Tanaka, K.; Funatsu, K.; Matsuura, F.; Soga, T.; Taguchi, R.; Saito, K.; Nishioka, T. *J. Mass Spectrom.* **2010**, 45, 703–714.

(40) Hummel, J.; Strehmel, N.; Selbig, J.; Walther, D.; Kopka, J. *Metabolomics.* **2010**, 6, 322–333.

(41) Hummel, J.; Selbig, J.; Walther, D.; Kopka, J. *Metabolomics.* **2007**, 18, 75–95.

(42) 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, 801–807.

(43) Stein, S. E.; Donald R. S. *J. Am. Soc. Mass Spectrom.* **1994**, 5, 859–866.

(44) Ibáñez, L.; Ong, K. K.; López-Bermejo, A.; Dunger, D. B.; de Zegher, F. *Nat. Rev. Endocrinol.* **2014**, 10, 499–508.

(45) Franks, S. *N. Engl. J. Med.* **1995**, 333, 853–861.

(46) March, W. A.; Moore, V. M.; Willson, K. J.; Phillips, D. I.; Norman, R. J.; Davies, M. J. *Hum. Reprod.* **2010**, 25, 544–551.

(47) Gambineri, A.; Patton, L.; Altieri, P.; Pagotto, U.; Pizzi, C.; Manzoli, L.; Pasquali, R. *Diabetes.* **2012**, 61, 2369–2374.

(48) Talbott, E. O.; Guzick, D. S.; Sutton-Tyrrell, K.; McHugh-Pemu, K. P.; Zborowski, J. V.; Remsberg, K. E.; Kuller, L. H. *Arterioscler. Thromb. Vasc. Biol.* **2011**, 20, 2414–2421.

(49) Tautenhahn, R.; Bottcher, C.; Neumann, S. *BMC Bioinformatics.* **2008**, 9, 1–16.

(50) Gómez, J.; Brezmes, J.; Rodriguez, M. A; Vinaixa, M.; Salek, R. M.; Correig, X.; Canellas, N. *Anal. Bioanal. Chem.* **2014**, 406, 7967–7976.

(51) Gómez, J.; Vinaixa, M.; Rodriguez , M. A; Salek, R. M.; Correig, X.; Canellas, N. *Advances in Intelligent Systems and Computing. 9th International Conference on Practical Applications of Computational Biology and Bioinformatics.* **2015**, 375, 59–67.

(52) Zhao Y.; Li Fu, L.; Li R.;, Wang L.; Yang Y.; Liu N.; Zhang C.; Wang Y.; Liu P.; Tu B.; Zhang X.; Qiao J. *BMC Medicine*. **2012**, 10, 1–12.

(53) Fang, M.; Ivanisevic, J.; Benton, H. P.; Johnson, C. H.; Patti, G. J.; Hoang, L. T.; Uritboonthai, W.; Kurczy, M. E.; Siuzdak, G. *Anal. Chem.* **2015**, 87, 10935–10941.

Table 1: Retention time (RT), quantitative fragment ion (m/z) (XCMS and MetAlign), relative concentration (Rel. C) and identification match factor (MF) (eRah) of 33 compounds. The table shows the coefficient of determination (R^2) of the regression between the area and intensity of the deconvolved compound elution profile (eRah) and the quantitative ion peak by XCMS and MetAlign (MetA). Percentage of variation between HIAE and control groups is also indicated for both compound (eRah) and peak (XCMS) intensity and area. The percentage was calculated as 100*(mean(HIAE)-mean(CTR)/mean(CTR)). The number of trimethylsilyl (TMS) derivative groups is shown in brackets.

						R^2			Percentage of variation (%)			
Cp.	Rt	m/z	Rel.C	Name	MF	XCMS		MetA	Area		Intensity	
No.#	(min)		(%)		(%)	Area	Int	Area	eRah	XCMS	eRah	XCMS
1	5.73	117	635	Lactic acid (2)	96.5	1.00	1.00	1.00	35	38	37	35
2	5.85	173	5	Hexanoic acid (1)	92.3	0.93	1.00	0.94	26	26	25	25
3	6.11	72	38	Valine (1)	96.8	0.97	1.00	0.98	18	18	21	21
4	6.53	249	63	Hydroxylamine (3)	96.4	0.82	0.91	0.78	-3	-6	-6	-7
5	6.69	131	11	2-hydroxybutyric acid (2)	98.2	1.00	1.00	1.00	2	4	1	3
6	7.08	86	20	Leucine (1)	99.1	0.98	1.00	0.97	23	23	25	25
7	7.15	191	11	3-hydroxybutyric acid (2)	90.8	1.00	1.00	0.99	-13	-13	-13	-12
8	7.97	145	20	Valine (2)	98.0	0.99	0.97	0.97	65	65	66	55
9	8.19	130	948	Urea (2)	91.9	0.93	0.92	0.90	13	19	15	17
10	8.36	179	76	Benzoic acid (1)	90.6	0.99	1.00	0.99	20	19	20	20
11	8.55	116	14	Serine (2)	95.0	1.00	1.00	0.99	48	46	49	48
12	8.75	158	11	Leucine (2)	98.6	1.00	1.00	1.00	85	73	77	70
13	8.80	205	122	Glycerol (3)	90.1	1.00	1.00	1.00	-8	-10	-7	-8
14	8.81	299	354	Phosphoric acid (3)	95.0	0.99	1.00	0.99	25	31	31	31
15	9.06	218	7	Isoleucine (2)	98.5	0.98	0.97	0.84	61	64	64	60
16	9.10	142	4	Proline (2)	97.8	1.00	1.00	0.99	56	57	66	63
17	9.58	189	3	Glyceric acid (3)	80.5	0.97	0.99	0.98	23	21	23	20
18	9.85	215	8	Nonanoic acid (1)	93.2	0.98	1.00	0.98	11	11	11	12
19	10.34	291	10	Threonine (3)	85.8	0.99	0.98	0.98	28	29	30	30
20	11.82	230	2	3-hydroxy-proline (3)	95.6	0.98	1.00	0.98	42	39	47	43
21	12.02	156	125	5-oxoproline (2)	99.5	1.00	1.00	1.00	35	36	35	34
22	12.73	142	7	Proline [+CO2] (2)	98.4	1.00	1.00	0.99	28	29	27	28
23	13.17	246	3	Glutamic acid (3)	92.0	0.99	1.00	0.99	118	105	107	103
24	13.27	218	14	Phenylalanine (2)	95.4	1.00	1.00	1.00	30	28	27	27
25	13.42	117	59	Dodecanoic acid (1)	92.4	0.91	1.00	0.93	5	6	3	3
26	15.40	142	9	Ornithine (4)	98.2	1.00	1.00	1.00	64	63	65	64
27	15.46	273	8	Citric acid (4)	96.7	1.00	1.00	1.00	25	24	26	25
28	15.54	285	20	Tetradecanoic acid (1)	91.7	0.99	0.94	0.95	4	6	2	5
29	16.43	230	15	Lysine (4)	94.3	0.35	0.84	0.45	45	30	28	31
30	17.48	129	414	Hexadecanoic acid (1)	91.1	1.00	1.00	1.00	7	6	8	6
31	18.23	305	17	Myo-inositol (6)	80.6	0.98	1.00	0.97	18	22	16	21
32	18.24	441	6	Uric acid (4)	92.0	0.99	1.00	0.98	2	6	4	6
33	19.26	356	224	Octadecanoic acid (1)	89.9	1.00	0.99	0.99	6	6	7	9

Table 2: Percentages of variation for lactate, urea, ornithine and myo-inositol using peak intensity (int) and area determined using eRah, XCMS, MetAlign (MetA), MassHunter (MH) and GC-QqQ MS analysis. The percentage was calculated as 100*(mean(HIAE)-mean(CTR)/mean(CTR)).

			QqQ	MH	MetA	eRah		XCMS	
Rt	Name	m/z	area	area	area	area	int	area	int
5.73	Lactic acid	117	32	39	39	35	38	38	35
8.19	Urea	130	1	13	19	15	13	19	17
15.40	Ornithine	142	44	64	67	64	65	63	64
18.23	Myo-inositol	305	11	23	23	18	16	22	21

Table 3: Percentage of variation and p-values (Wilcoxon–Mann–Whitney test) of statistically significant metabolites. The positive variations indicate higher levels in girls with HIAE relative to healthy controls.

Rt	Name	p-value	%Var
5.73	Lactic acid (2TMS)	0.0090	35
7.08	Leucine (1TMS)	0.0014	23
8.55	Serine (2TMS)	0.0022	48
8.75	Leucine (2TMS)	0.0034	85
11.82	5-oxoproline (2TMS)	0.0042	35
13.17	Glutamic acid (3TMS)	0.0028	118
15.40	Ornithine (4TMS)	0.0002	64
16.43	Lysine (4TMS)	0.0034	45

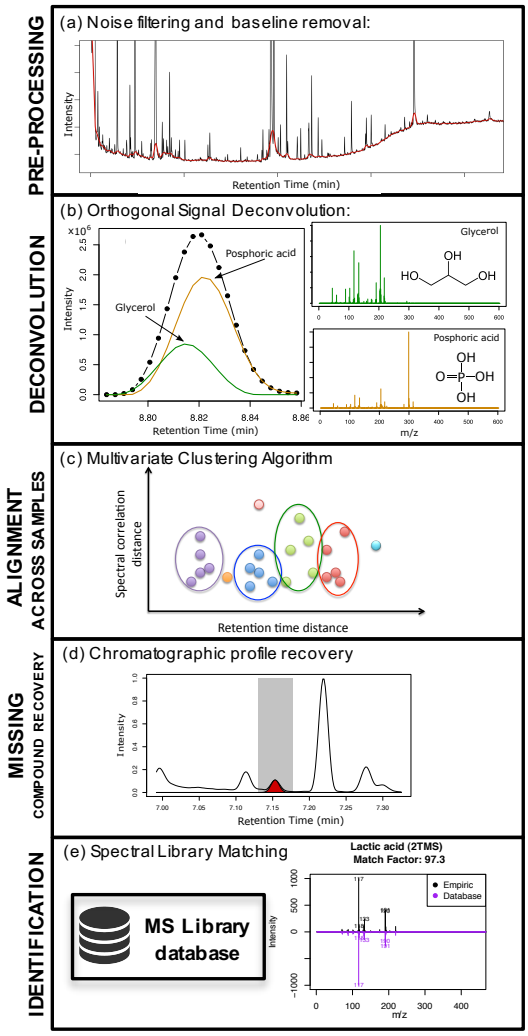


Figure 1: eRah's workflow. First, a pre-processing step (a) is applied to remove the noise and the baseline (red) from the chromatogram (black). Second, the deconvolution stage (b) extracts the chromatographic compound profiles and spectra from each sample. Third, compound spectra are aligned (c) across all samples and a missing compounds recovery step (d) retrieves those compounds that were not found in certain samples. Finally, extracted spectra are matched against an MS library (e), providing a list of metabolites and their intensity (or area) in each sample.

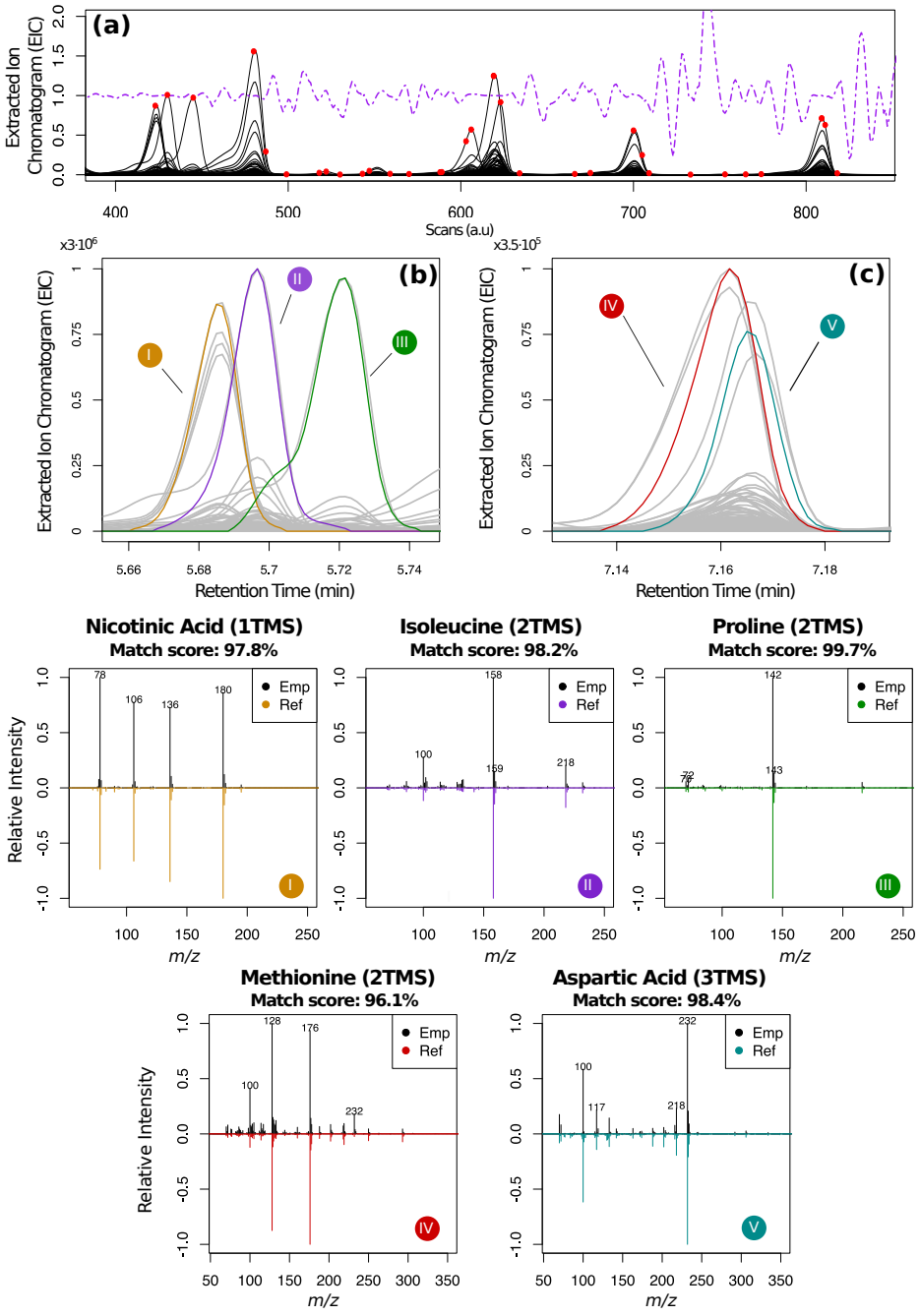


Figure 2: Top image (A), shows the operation of the CMLC filter: the black lines depict extracted ion chromatograms (EIC) in the sample, the purple line is the filter output characterized by local minima (marked with red dots in the EIC). Figures B and C show two co-elution situations. The extracted ion chromatograms are shown, where each gray line corresponds to a different m/z peak. Colored solid lines are the deconvolved profiles of the compounds. The deconvolved spectra for each compound are shown in black in figures along with each reference spectrum negatively rotated in the same axis. The match factor is also noted.

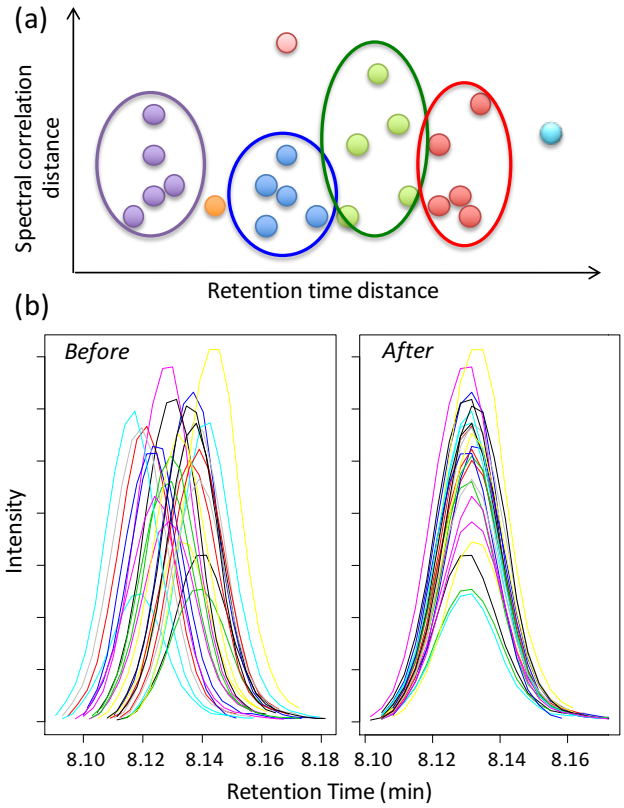


Figure 3: (a) Representation of the alignment algorithm. The spheres represent four resolved compounds after deconvolution by eRah. Each compound (purple, blue, green and red spheres) appears in five different samples. We have included three additional compounds as an interference (orange, pink and light blue sphere). The compounds are projected into a two-dimensional space for illustration purposes where their proximity is determined by the spectral similarity and retention time distance. The algorithm aims to cluster the same compound in one group on the basis of proximity in spectra similarity and retention time. (b) Elution profile of urea across samples before and after alignment.

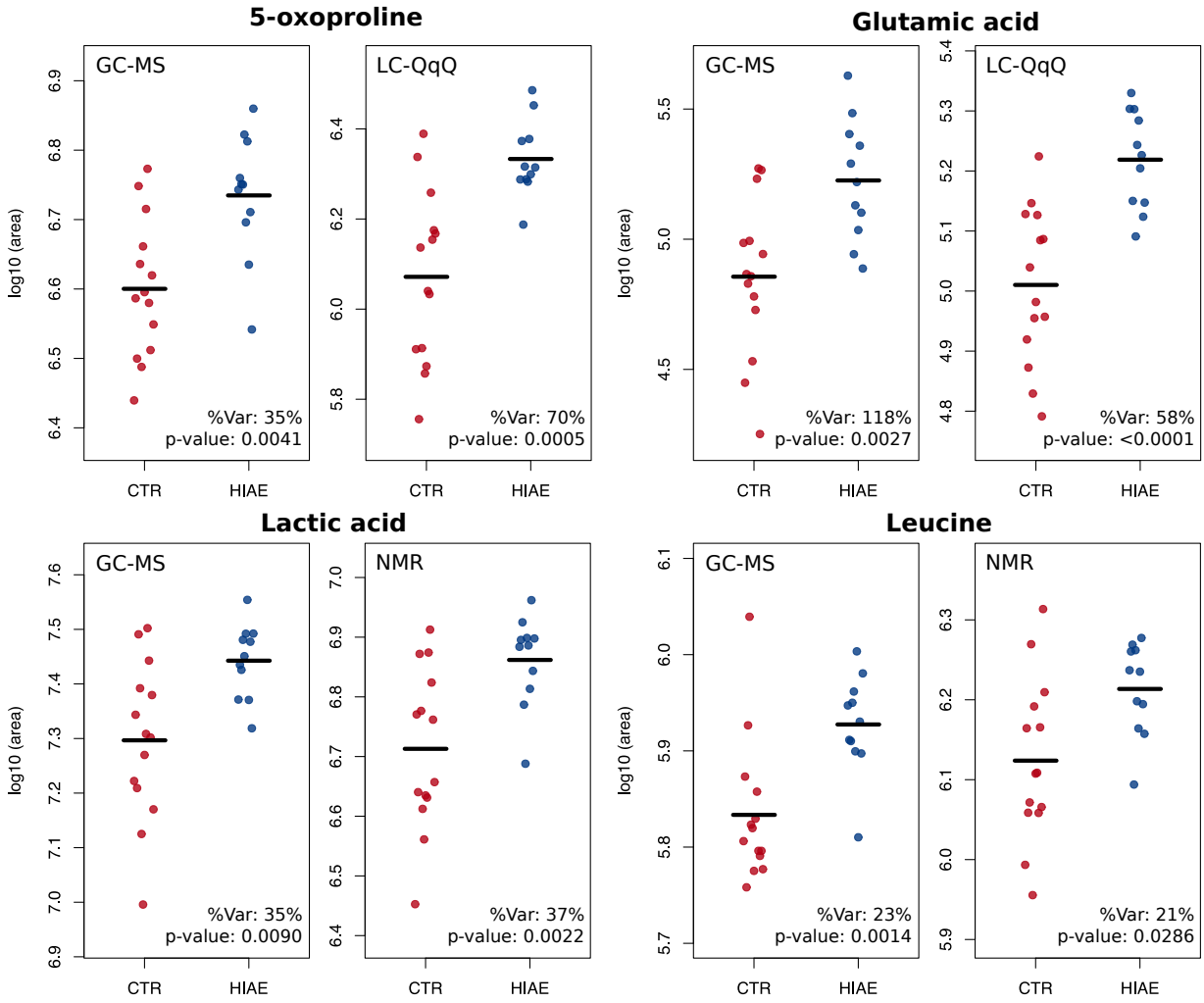
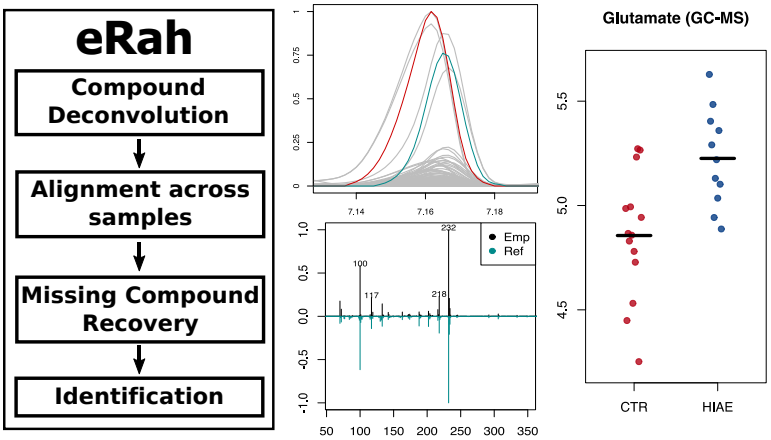


Figure 4: Scatter plots of metabolites identified and quantified by GC–MS (eRah), LC–QqQ-MS targeted analysis and NMR. The scatter plots show the abundance of 5-oxoproline, glutamic acid, lactic acid and leucine in controls and HIAE serum samples and trimmed mean (controls are depicted in red and HIAE in blue). Percentage variation (%Var) and p-values (Wilcoxon–Mann–Whitney test) are also shown.

TOC entry:


eRah
Compound Deconvolution
Alignment across samples
Missing Compound Recovery
Identification
Emp
Ref
100
117
218
232
Glutamate (GC-MS)
CTR
HIAE