Journal of MASS SPECTROMETRY

Received: 17 June 2014 Revised: 13 September 2014 Accepted: 17 September 2014 Published online in Wiley Online Library

(wileyonlinelibrary.com) DOI 10.1002/jms.3512

GridMass: a fast two-dimensional feature detection method for LC/MS

Victor Treviño,[a]* Irma-Luz Yañez-Garza,[a] Carlos E. Rodriguez-López,[b] Rafael Urrea-López,[c] Maria-Lourdes Garza-Rodriguez,[d] Hugo-Alberto Barrera-Saldaña,[d] José G. Tamez-Peña,[a] Robert Winkler[e] and Rocío-Isabel Díaz de-la-Garza[b]

One of the initial and critical procedures for the analysis of metabolomics data using liquid chromatography and mass spectrometry is feature detection. Feature detection is the process to detect boundaries of the mass surface from raw data. It consists of detected abundances arranged in a two-dimensional (2D) matrix of mass/charge and elution time. MZmine 2 is one of the leading software environments that provide a full analysis pipeline for these data. However, the feature detection algorithms provided in MZmine 2 are based mainly on the analysis of one-dimension at a time.

We propose GridMass, an efficient algorithm for 2D feature detection. The algorithm is based on landing probes across the chromatographic space that are moved to find local maxima providing accurate boundary estimations. We tested GridMass on a controlled marker experiment, on plasma samples, on plant fruits, and in a proteome sample. Compared with other algorithms, GridMass is faster and may achieve comparable or better sensitivity and specificity.

As a proof of concept, GridMass has been implemented in Java under the MZmine 2 environment and is available at http://bioinformatica.mty.itesm.mx/GridMass and MASSyPup. It has also been submitted to the MZmine 2 developing community.

Additional supporting information may be found in the online version of this article at the publisher's web site.

Keywords: metabolomics; feature detection; software & algorithms; HPLC/MS; MZMine 2

Introduction

The wide and unbiased analysis of metabolites in metabolomics is an important technique to study organisms, tissues, and cells. For this, high-performance liquid chromatography coupled to mass spectrometry (HPLC/MS) is one of the most applied technologies.[1] The output of HPLC/MS experiments is a three-dimensional matrix representing molecular mass, time, and detected abundance. It requires sophisticated computational analyses to extract meaningful information. Some pipelines have been implemented in free and open software packages such as MZmine 2, XCMS, MaxQuant, and OpenMS[1−5] including installation-free distributions such as MASSyPup.[6] The complete analysis pipeline involves a mandatory feature detection step. Feature detection is the procedure to detect the boundaries of a putative molecule within the mass and time domains. Some methods have been proposed for this task.[2,3,7,8] Detection of high intensity features can be easily achieved because the signal to noise is high. The detection of low-intensity masses is nevertheless more challenging and often yields false positives. It has been proposed that two-dimensional (2D) feature detection may be more efficient than methods that perform a two-step peak detection such as centroid picker in MZmine 2.[3] Nevertheless, there are no available 2D feature detection modules for MZmine 2, which is one of the most user-friendly software packages for metabolomics analysis having many user advantages such as a multiplatform and simple operation by non-expert users.

For these reasons, we propose GridMass, a highly sensitive feature detection algorithm. As a proof of concept and to facilitate usage by non-expert users, we implemented GridMass for MZmine 2. We provide a clear pipeline to use GridMass. Further, we show its utility on three types of experimental data and compared the results with other algorithms. In addition, Supporting Information shows a simple proteomics example. Our results demonstrate that GridMass is a useful tool for HPLC/MS analysis within the MZmine 2 framework.

* *Correspondence to: Victor Treviño, Cátedra de Bioinformática, Departamento de Investigación e Innovación, Escuela de Medicina, Tecnológico de Monterrey, Guadalupe, Nuevo Leon, 64849, México. E-mail: vtrevino@itesm.mx*

a *Cátedra de Bioinformática, Departamento de Investigación e Innovación, Escuela de Medicina, Tecnológico de Monterrey, Guadalupe, Nuevo Leon, 64849, Mexico*

b *Cátedra de Micronutrientes, Escuela de Biotecnología y Alimentos, Centro de Biotecnología - FEMSA, Tecnológico de Monterrey, Monterrey, Nuevo Leon, 64849, Mexico*

c *Cátedra Empresarial de Tecnologías de Agricultura Intensiva, Escuela de Biotecnología y Alimentos, Centro de Biotecnología - FEMSA, Tecnológico de Monterrey, Monterrey, Nuevo Leon, 64849, Mexico*

d *Departamento de Bioquímica y Medicina Molecular, Facultad de Medicina, Universidad Autónoma de Nuevo León, Monterrey, Nuevo Leon, 64460, Mexico*

e *Departamento of Biotecnología y Bioquímica, CINVESTAV Unidad Irapuato, Irapuato, Guanajuato, Mexico*

Methods

The GridMass algorithm

High-performance liquid chromatography coupled to mass spectrometry data can be represented by a 2D image displaying time and molecular weight in *x*-axis and *y*-axis, respectively, where the color represents intensity (amount of mass detected) such as in Fig. 1(A). To detect the position and boundaries of masses, the GridMass algorithm first generates a grid of equally spaced probes covering the entire chromatographic area. A representative section is shown in Fig. 1(B). Each probe [black dots in Fig. 1(B)] explores a rectangular region around it to find a local maximum. The probe location is moved to the local maxima to further search for a higher value. The procedure is performed until no higher values exist within the exploring rectangle. This local maximum is then defined as a *feature*, which contains information of the *m*/*z*, the time, and the intensity detected. A putative trajectory of a probe is shown in Fig. 1(B). All probes converging to the same feature provide an estimation of its boundaries. Consequently, different features represent different masses. This procedure is highly sensitive and specific for smooth surfaces. However, given that real chromatographic data display a certain level of noise and many other artifacts, additional criteria were implemented. A summary of the considered parameters is shown in Table 1. The complete algorithm (highlighting the parameters in bold and italics) has been implemented in the following steps:

(1) Ignore artifact spectra in time domain. In chromatography, it is typical to find a peak near the injection time, corresponding to metabolites that show no interaction with the column in the particular gradient. Although this is not an artifact *per se*, given the myriad of signals present and the nature of the detector, the resulting peak is a strong source of artifacts that later affect analysis. An example of such unspecific elution is seen in Fig. 1(A) around time = 2.5. This issue may generate many false features. To avoid this, the user may enter a list of time ranges to be ignored. The controlling parameter is ***ignore times*** whose format is time1-time2, time3-time4,…; alternatively, the user may crop these data before processing, for example, using 'Raw Data→Filtering→Data Set…' in the MZmine 2 main menu and ignore setting this parameter. Therefore, this step is optional.

(2) Generate equally spaced probes over the mass-time space. To generate the grid, the parameters used are ***m*/*z* tolerance,** and ***minimum width***. The gap in the *m*/*z* dimension between probes is set to *m*/*z tolerance* multiplied by 2 or minimum to 1e^{-6} and are intercalated between scans. The gap in the time dimension is calculated by time associated to scans, which is estimated by the ***minimum width*** divided by 4 (down to a minimum of one scan).

(3) Move each probe to the corresponding local maximum until convergence. As shown in Fig. 1(B), each probe explores its surroundings (limited by the positions of other probes) to locate the highest intensity value, then after updating its position, it keeps exploring the surrounding until a local maximum is reached. To speed up the procedure, generate only interesting features above a certain level of noise, and limit the number of reported features, only intensities higher than ***minimum height threshold*** are considered.

(4) Generate features by merging probes with similar 2D positions. Many probes will reach the same maximum that must correspond to the same feature. In addition, experimental chromatographic data are noisy and non-smooth, which may generate local maxima very close to each other. Therefore, probes whose difference in *m*/*z* is lower than the ***m*/*z tolerance*** and whose difference in time is lower than ***minimum width*** are merged. Then, from all probes reaching the same maximum, the *m*/*z* assigned to the feature corresponds to the highest observed intensity. The width of the feature is estimated from the probes with the lowest and highest time. To form the peak and estimate its area, the highest value in each scan is used.

(5) Remove features whose width is out of a range. Features having large or very low width are likely to be artifacts. To avoid this, all features out within the range given by the parameters ***minimum width*** and ***maximum width*** (in minutes) are removed.

(6) Remove features of similar mass and high cumulative times. Chemical noise or large blurs are characterized by generating many features of similar mass, similar intensities, and separated by short times. To avoid these artifacts, we merge features whose *m*/*z* difference is lower than ***m*/*z tolerance*** and whose intensity ratio (higher/lower) is higher than an ***intensity similarity ratio*** parameter. Once merged, the removal implemented in Step 5 is performed on merged features.

The GridMass algorithm was implemented in Java as a module for the MZmine 2, which can be downloaded from http://bioinformatica.mty.itesm.mx/GridMass. It was also included within the MASSyPup distribution[6] and has also been submitted to the MZmine 2 developing community, which could be included in future versions of MZmine 2 facilitating the forthcoming availability.

Supporting Information shows precise instructions on how to install and configure computer memory and step-by-step instructions of how parameters can be estimated.

To evaluate the performance of GridMass, we performed three experiments that represent typical assays using standards, blood, and plant samples recorded in centroid, continuous, and centroid modes, respectively. In addition, a proteomics example is briefly shown in the Supporting Information.

Experiment 1: MM14 standards

The aim of this experiment is to show the accuracy of the proposed GridMass algorithm. For this, we used a dataset originally presented with the centWave algorithm where a mixture of 14 markers has been annotated. The centWave algorithm is based on a wavelet algorithm to detect centers and border of peaks.[3] We therefore compared GridMass with the previously optimized parameters for centWave. We also tested other two algorithms. In addition, to briefly show the sensibility of GridMass to parameters, we made comparisons of parameter values.

Dataset

We analyzed the MM14 experiment containing a mixture of 14 marker compounds[3] in which the expected adduct masses have been annotated. The dataset was recorded in centroid mode. The annotation consists of the molecular formulae, the atomic mass, the retention time, and the mass percentage relative to maximum peak. In total, 296 compounds were annotated. However, 22

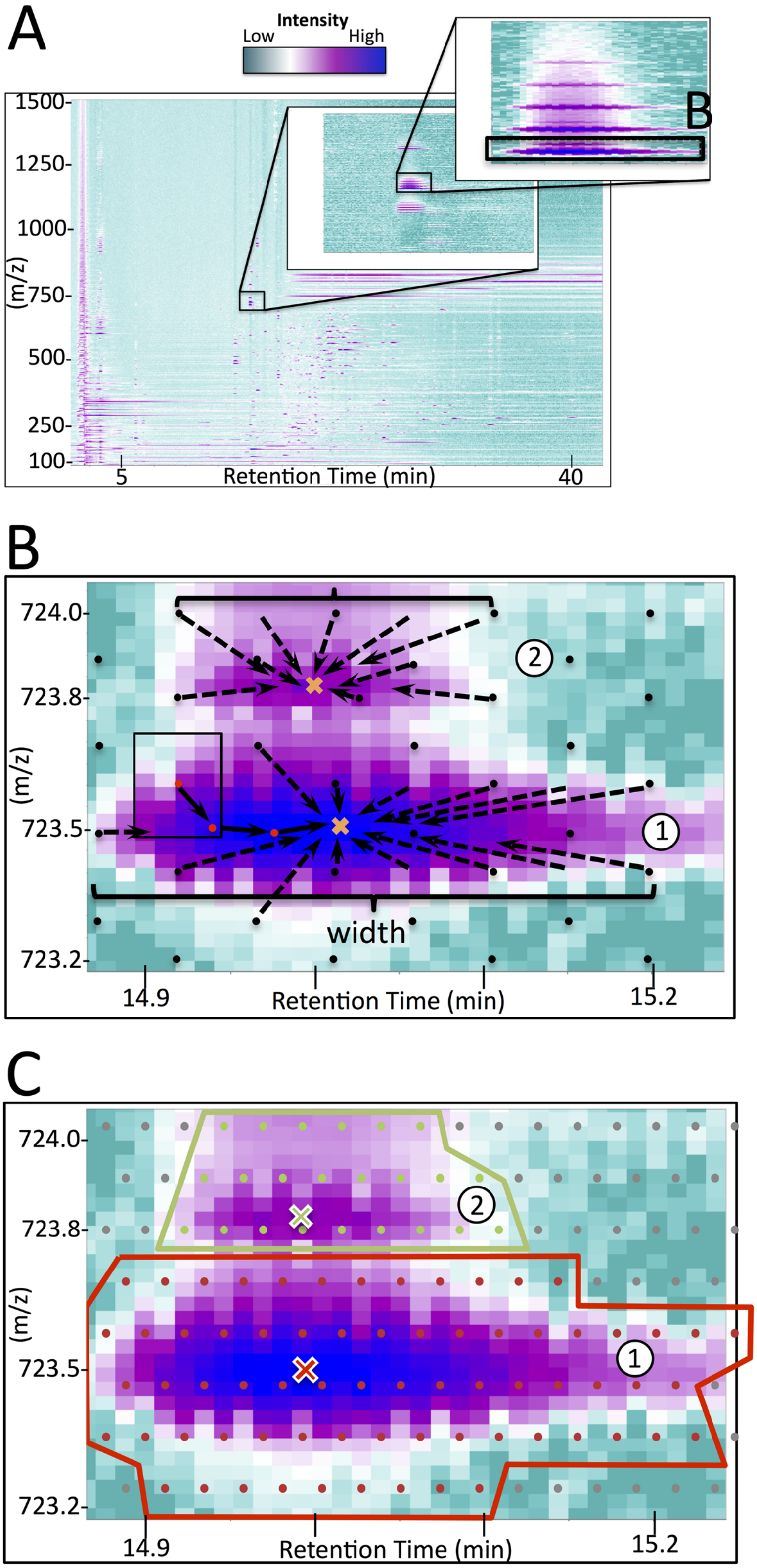

Figure 1. Graphical overview of the GridMass algorithm. Panel A shows an image representation of the whole chromatography showing zooms (*m/z* in vertical axis and retention time in minutes in horizontal axis). Dashed box is shown in panel B. Intensities are drawn in $\log_{10}$ scale. Panel B depicts the GridMass algorithm for two peaks. The probes are represented by black dots that will be moved to a local optimum. The continuous arrows represent the movement of a particular probe (in red) to a center marked with 'x'. Dashed arrows represent the overall movement of remaining probes. The explored area in each movement is delimited by a rectangle relative to the probe center and limited by neighbor probes or minimum values (2 scans and twice of the *m/z* tolerance). The colors in the 2D chromatogram represent intensity of detected masses, cyan, white, magenta, and blue for very low, low, medium, and high intensity values respectively. Panel C shows the actual detection of the same region in B from Experiment 2 using minimum height = 50 and no smoothing. Green and red dots represent probes assigned to peaks 1 and 2, respectively. Corresponding centers are shown as a colored 'x'. Colored polygons show the approximated boundary estimations.

Table 1. Summary of parameters, their use, and recommended values

Parameter	Use	Recommended value
minimum height threshold	Step 3: intensities lower than this value will be ignored.	Between 10 and 1000 depending on the background noise. See MS plot in a representative time.
width	Step 2: gap between probes = width/4. Steps 4 and 5: Min and Max time = width.	From 1 to few seconds depending on the chromatographic conditions. See 2D plot.
m/z tolerance	Step 2: gap between probes = *m/z* multiplied by 2. Steps 4 and 6: Maximum difference in *m/z* of two probes/features that can be considered the same.	Between 0.001 and 0.1 depending on the mass spectra resolution. See 2D plot.
intensity similarity ratio	Step 6: to detect artifact features having similar intensity and mass.	Between 0.1 and 0.5 depending on the observed artifacts. See 2D plot.
ignore times	Optional. Step 1: list of time ranges to be ignored.	Depends on the sample and chromatographic conditions. See 2D plot.
smoothing time	Optional, performed before Step 3: time considered for smoothing the chromatogram by averaging.	Time required for three to seven scans.
smoothing m/z	Optional, performed before Step 3: *m/z* range for smoothing the chromatogram by averaging.	Lower than ***m/z tolerance***. Depends on the MS precision.

compounds weighting less than 100 Daltons were removed because they fell below the limit of the mass detector. Thus, the final annotation contained 274 compounds.

Algorithms and parameters

To show the performance of GridMass, in this analysis, we compared its results against those using centWave.[3] To determine if an algorithm found a particular compound, we considered a feature as found when the difference between the annotation and the reported mass (in *m/z*) and retention time (in minutes) by the algorithm was lower than 0.1. We used the recommended parameters for centWave.[3] For GridMass, we used the same height threshold suggested by centWave. Other parameters for GridMass were estimated as shown in Fig. S1 (Supporting Information) and summarized in Table S1 (Supporting Information). We also tested Centroid in MZmine and OpenMS using equivalent parameters. To show the sensitivity and dependency of GridMass for parameters, we ran GridMass in various settings. Both algorithms were executed in a MacBookPro having OS X 10.8.4, 16 GB of RAM (1600 MHz DDR3) and 2.5 GHz Intel Core i5. The running time was estimated by representative runs.

Experiment 2: metabolomics profiles from plasma donors

In this example, we aimed to show how to use GridMass within a metabolomics analysis pipeline from an LC/MS experiment. For comparison, we performed the procedure for MZmine 2 with GridMass and for XCMS with centWave. Overall, the procedure performed is depicted in Fig. 2(A).

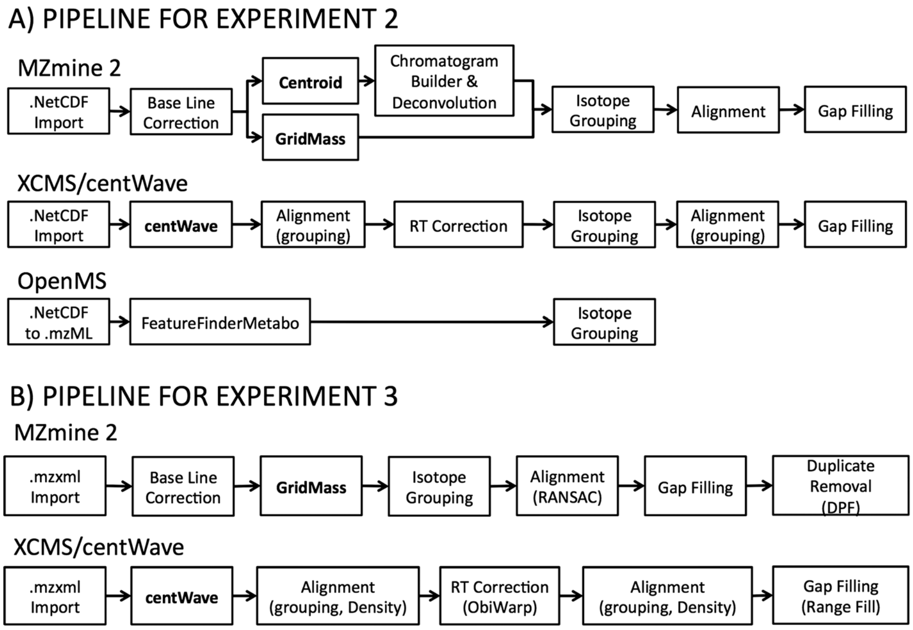

Figure 2. Data analysis protocol for experiments 2 and 3.

Plasma samples

We used a preliminary sample collection from presumably healthy women attending annual mammography screening at a clinic in northwest Mexico (Instituto Mexicano del Seguro Social, clinic 17) and whose radiological findings were normal. The institutional review board approved the protocol. After the proper informed consent acceptance, 6 ml of blood from 17 donors was collected in Vacutainer tubes containing 10.8 mg of Ethylenediaminetetraacetic acid (EDTA) and kept on ice for no more than 6 h. Plasma was obtained by centrifugation at 3000 g for 20 min at 4 °C and stored at −80 °C in a freezer in 350 µl aliquots until analyzed (less than 5 months). All samples were further processed in a single batch.

Metabolite isolation

Before the LC/MS analysis, the protein from plasma aliquots was removed by thawing the plasma aliquot on ice for 30 min, then adding 21.4 µl of 20.5 mg/ml umbeliferone (Sigma) as an internal reference and 450 µl of ethanol at −20 °C, centrifuging at 15 800 g, recovering 500 µl of the supernatant, filtering in a tube having a 0.22-µm nylon filter, centrifuging at 1100 g for 2 min, and making aliquots of 150 µl.

Chromatography and mass spectrum

The separation of a 15-µl sample was performed in an HPLC Agilent 1100 series equipped with a C18 column (Luna-C18 3 µm, 2.0 × 150 mm and Luna-C18 4.0 × 2.0 mm from Phenomenex) configured at a flow rate of 190 µl/min applying a nonlinear gradient of phase A (water +0.1% formic acid) and phase B (acetonitrile, 0.1% formic acid) as follows: 5–40% B (0–10 min), 40–75% B (10–13 min), 75–95% B (13–23 min), 95% B (23–36 min), 5% B (36–40 min), and 5% B (40–42 min). Eluted compounds were detected in the 100–1500 *m/z* range by an electrospray-ionization time of flight device (Agilent G1969A) in negative ion mode using the following settings: nebulizer N_2 gas pressure, 40 psig; drying gas, N_2 at 300 °C, 11 l/min; capillary 4000 V; fragmentator 175 V; skimmer 60 V; and OCT RFV 250 V. We used the vendor's software for data acquisition and recorded as profile mode (continuous) for 41.9 minutes registering 3652 scans (1.45 scans/s).

Data processing

Raw data (.wiff) were converted to NetCDF format using the File Translation Utility software from Agilent. The average file size was 560 MB. An in-house build of MZmine 2 that includes the GridMass module was used, which is deposited in http://bioinformatica.mty.itesm.mx/GridMass.

Baseline correction

In MZmine 2, the imported files were baseline corrected using the option *Baseline correction* and the following parameters: chromatogram type = base peak intensity; MS level = 1; smoothing = 10 000; asymmetry = 0.001; use *m/z* bins checked; and *m/z* bin width = 1.

Feature detection

For GridMass, we estimated the parameters from the 2D chromatographic images in MZmine 2. The settings were as follows *m/z* tolerance = 0.10; minimum width = 0.05; maximum width = 5; smoothing time (min) = {0, 0.05}; smoothing *m/z* = {0, 0.05}; false+: intensity similarity = 0.5; false+: ignore times = 0-5; minimum height was varied from 100 to 10, specifically to {100, 50, 25, 20, 15, and 10}. Tests were performed in a specific sample to estimate the best parameters to use. Then, all 17 samples were processed using the chosen parameters. For the equivalent peak detection procedure in MZmine 2 not using GridMass, we used the default packages and mainly varied the height. We used the *Centroid* detection, setting noise levels on different runs to one of {100, 50, 25, 20, and 10}. Correspondingly, we used *Chromatogram Builder* setting minimum height, respectively, to {100, 50, 25, 20, and 10}; minimum time span = 0.05; *m/z* tolerance = 0.05 and 0 ppm. Finally, we used *Chromatogram deconvolution* using the *local minimum search algorithm* setting chromatographic threshold = 70%; search minimum in retention time (RT) range = 0.05; minimum relative height, respectively, to {100, 50, 25, 20, and 10}; minimum ratio of peak top/edge = 1.5; peak duration time = 0.05 to 5.0. For centWave, we used the parameters shown in Table S2. For OpenMS, we used MASSyPup[6] and the parameters shown in Table S2.

Isotope grouping

Features that follow an isotopic pattern were grouped in MZmine 2 using the option *Isotope Peak Grouper* and the following parameters: *m/z* tolerance = 0.02 and 0 ppm; RT tolerance = 0.1 absolute; monotonic shape = unchecked; maximum charge = 1; representative isotope = most intense. To avoid duplicated isotope patterns, isotopes with a mass difference less than 0.15 and whose RT difference was less than 10 s were considered as the same peak, removing the one with lower median area using the option *Duplicate peak filter*. Non-isotopes were removed by using the option *Peak list row filter* setting minimum peaks in an isotope pattern = 2. For centWave, we implemented in R an algorithm to find putative isotopes. For each mass, we looked up similar masses adding from 1 to 6 and whose maximum mass difference was 0.15 and whose RT difference was less than 10 s. Features having at least one isotope were considered a positive isotopic pattern. To avoid duplicated isotope patterns, isotopes of mass difference less than 0.15 and whose RT difference was less than 30 s were merged.

Alignment

To correct differences in time and masses between samples, the features were aligned in MZmine 2 using the *RANSAC aligner* algorithm and the following parameters: *m/z* tolerance = 0.1, RT tolerance = 1.0 min, RT tolerance after correction = 1.1 min, RANSAC iterations = 100, minimum number of peaks = 30%, threshold value = 1.5, linear model = checked.

Quality measurements

We counted the number of features having an isotopic pattern as a sensitivity measure of the algorithms. Nevertheless, solvents and many artifacts and chemical noise may also present isotopic patterns, which may obscure conclusions. For this, we avoided noisy chromatographic regions; therefore, we performed the feature detection in the whole data but measured quality in a 'relatively clean' region for masses between 350 and 700 *m/z* and time between 16 and 25 min as shown in Fig. S2. The counting was performed in a representative sample (C6).

Experiment 3: metabolomic profiling of habanero pepper samples

To challenge GridMass with more complex samples, we also performed a controlled experiment in plants. Using GridMass and CentWave, we analyzed seven biological replicates of Habanero fruit along with three technical replicates injected non-consecutively. To minimize external influence from the measurement procedure, a repeated measurement ANOVA (rANOVA) was preferred given its capability to partition and isolate variability from the repetitions. The pipeline used is shown in Fig. 2(B). We used typical working parameters for both procedures trying to maintain similar values whenever possible.

Plant material

Habanero peppers (*Capsicum chinense*, Jacq.) of the 'Orange' variety (Seminis Inc. St. Louis, MO, USA) were grown in perlite under controlled conditions in a greenhouse at the Tecnológico de Monterrey in Monterrey, Mexico (25°40′N 100°18′W, altitude 430 m). Plants were irrigated daily with water and twice a week with Hoagland solution.[9] For this study, seven plants were selected, from which ripe fruits were harvested 3 weeks after the onset of ripening (breaker stage). Pericarp tissue was separated, frozen in liquid nitrogen, and preserved at −80 °C until metabolite analysis.

Sample preparation

Each replicate consists of a pool of six fruits from the same plant, which were homogenized under liquid nitrogen in a frozen mortar. Approximately 0.5 g of frozen pericarp powder was extracted according to de Vos[1] in methanol with 0.125% formic acid, sonicated for 15 min and then separated by centrifugation (14 000 g at 4 °C) for 10 min. The supernatant was filtered using syringe filters (0.2 μm) and collected in amber glass vials. Extracts were separated via HPLC using a Luna C18(2) column (Phenomenex, cat.00 F-4251-B0, Torrance, CA, USA) maintained at 40 °C, with a 60 min nonlinear gradient of water/formic acid 0.1% (phase A) and acetonitrile/formic acid 0.1% (phase B), and measured with the aid of a coupled time of flight mass spectrometer (Agilent LC/MSD TOF, G1969A, Santa Clara, CA, USA). Data were acquired using the vendor's software and recorded in centroid mode in a range from 80 to 1800 *m/z*.

Data processing and statistical analysis

Raw data (*.wiff) were converted to netCDF (*.cdf) format using the Translation Utility software from Agilent. The same files were analyzed by MZmine 2 with the in-house built GridMass algorithm and with XCMS within R. Between algorithms, features were considered to be the same if their difference in average *m/z* was less than 50 ppm and less than 30 s in mean RT or if its average *m/z* and RT were within the boundaries of the minimum and maximum of the other. Whenever a conflict was encountered (i.e., there was more than one feature with those characteristics), a correlation matrix was generated, and the pair corresponding to the highest correlation was conserved. rANOVA was performed using the 'ez' package in R[10] considering plants as individuals and injections and repetitions. A feature was considered a false positive if the effect of the repetition was significant with a p-value < 0.01 or p-value < 0.001 for comparisons.

Feature detection

Peak detection was performed on MZmine 2 with the GridMass algorithm with the following parameters: minimum height = 1000 or 100; *m/z* tolerance = 0.1; minimum width = 0.167; maximum width = 1.5; smoothing time (min) = 0.15; smoothing *m/z* = 0.07; False+: intensity similarity ratio = 0.5; false+: ignore times = 0–5. The parameters for detection using CentWave on XCMS were as follows: step = 0.1; ppm = 30; peak width = (10, 90); snthresh = 6; prefilter = (3, 1000) or (3, 100); mzdiff = 0.1; and noise = 100.

Isotope grouping

Isotope filtering was performed in both MZmine and XCMS results using an in-house built program, as mentioned in Example 2. Namely, the files were ordered by *m/z,* and, row by row, the corresponding isotopes (from M + 1 to M + 6, single charged) were searched among the detected features, allowing a tolerance of 0.15 *m/z* absolute value and 10 s of RT. Only features for which at least one isotope was found were kept, saving only the monoisotopic moiety.

Alignment

Alignment was executed on MZmine using the RANSAC aligner with a tolerance of 0.05 *m/z* or 30 ppm, an RT tolerance of 1.5 min and 1.0 min after correction, 2000 iterations, a minimum number of points set to 30%, and a threshold value of 0.05, using a linear model. On XCMS, it was performed by using the density grouping (which is the default) with the following parameters: bw = 5, minfrac = 0.5, minsamp = 1, mzwid = 0.015, and max = 100.

RT correction

Retention time correction was performed as a separate process on XCMS, as it is already integrated in the RANSAC on MZmine 2. We used ObiWarp ran with a profStep = 1.

Gap filling

A gap filling was performed on both pipelines on features whose intensity was zero, by integrating the intensity in the *m/z* values within an RT window in case a peak was not detected in a sample.

Duplicate peak filtering

Finally, exclusively on MZmine, peaks considered to be the same feature were removed by using a duplicate peak filtering. Peaks whose *m/z* difference was less than 0.05 or 30 ppm, with an RT tolerance of 30 s, were removed.

Results

Figure 1(C) shows the proof of concept experiment on a representative example of two detected features and its estimated boundaries. The figure shows the grid of probes together with those that moved to the reported features center and the estimated boundaries. It is clear that the algorithm successfully discriminated between neighboring masses and provides good approximations of the overall feature surfaces.

The following sections will provide detailed performance and comparison results on the three experiments performed.

Experiment 1: MM14 standards

We analyzed the MM14 experiment containing a mixture of 14 marker compounds[3] in which the expected adduct masses have been annotated. The final used annotations contain 274 compounds. An overview of this data and GridMass parameter estimation is shown in Fig. S1 (Supporting Information). As shown in Table 2, the GridMass algorithm performed slightly better than centWave detecting two more true features and one less false positive. In addition, GridMass took 3 s less CPU time. Centroid and OpenMS showed lower performance. Although the differences in the results for this dataset are minor, they suggest that GridMass is competitive and useful.

Table 2. Features detected by two algorithms depending on parameters

Algorithm	Detected MM14 features	Putative false positives	Elapsed time (s)	Parameters used
centWave	140	887	8	Peak width = (5 s, 10 s), prefilter = (2, 200), ppm = 30, snthresh = 4
GridMass*	142	886	5	Width = (0.02 min, 0.50 min), minimum height = 200, *m/z* tolerance = 0.05, similarity intensity = 0.5
Centroid	102	758	180	Centroid mass det. level = 200, chr. builder time = 0, height = 200, *m/z* tol = 0.05; Loc. min. search deconv., thr = 70%, RT = 0, rel. height % = 0, abs. height = 150, ratio = 1, duration = 0 ~ 5
OpenMS	47	133	4	noise thr. int = 200, chrom peak snr = 4, chrom fwhm = 10 mass error ppm = 30, reest. mt sd = true, width filt. = auto, charge = 1– 3

* Optional parameters were set to 0: no smoothing (time = 0, *m/z* = 0) and ignore times = 0–0.

We then analyzed the response of GridMass to changes in parameter values. We used the settings of first comparison as the baseline parameters and then changed one parameter at a time. The results shown in Table S1 (Supporting Information) suggest that, to detect more MM14 features, the *minimum height* is the most important parameter followed by *width*. The running time ranged from 5 to 14 s suggesting that changes in parameters do not compromise the running time in this data.

Overall, the MM14 experiment indicates that GridMass is sensitive, fast, and competitive, and its performance depends mainly on the minimum height parameter.

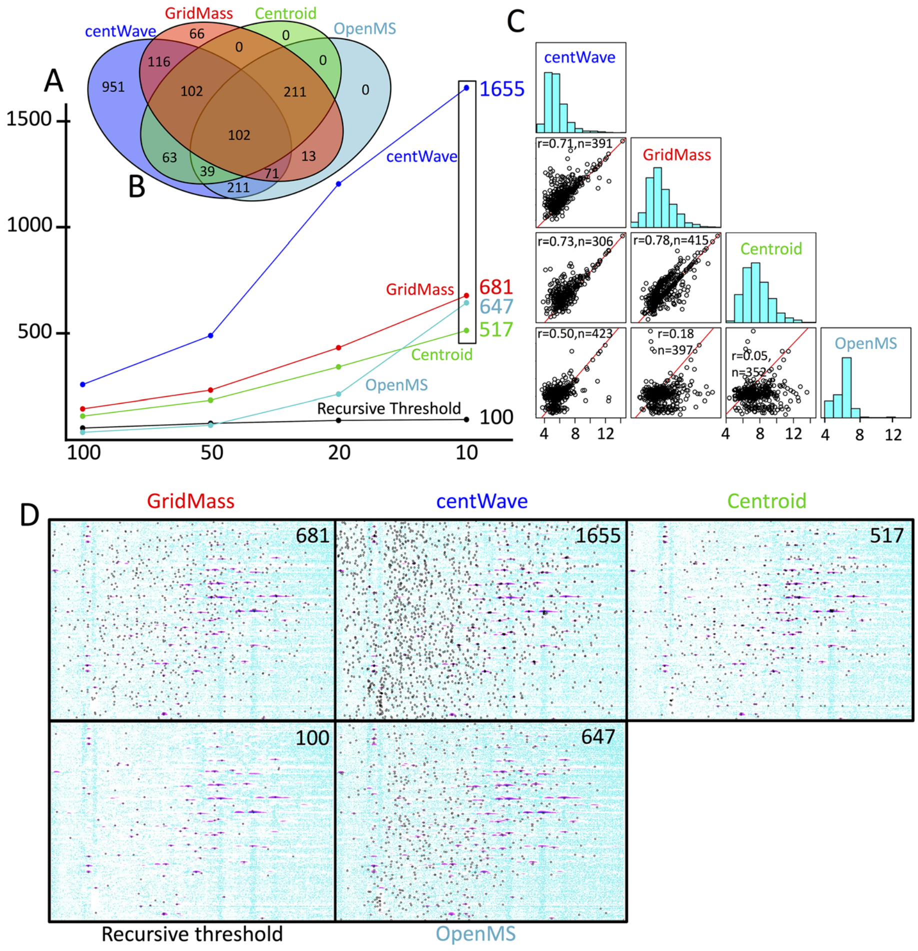

Figure 3. Summary of results from the plasma samples of experiment 2. Panel A shows the number of masses per method that follow an isotope pattern. Four representative height thresholds (100, 50, 20, and 10) are shown in the *x*-axis. Four methods detecting more than 500 masses at height = 10 (black rectangle) were used for comparisons in panels B and C. Panel B shows a Venn diagram of the number of similar masses per method. Masses were considered similar if their Euclidean distance of *m/z* and retention time were lower than 0.1. Panel C shows that log-scaled estimates of the peak area of similar masses across methods are highly correlated using Pearson correlation coefficients except for OpenMS. Panel D shows a section of the 2D chromatogram comparing the detected masses (black dots) per method using height = 10. The horizontal axis shows retention time from 16 to 25 min, whereas the vertical axis displays *m/z* from 350 to 700.

Experiment 2: metabolomics profiles from plasma donors

One of the main objectives of feature detection methods is to detect as many true features (masses) and less false features as possible. We hypothesized that true masses should be distinguished by an isotopic pattern. Therefore, to challenge sensitivity in a typical non-replicative setting, we compared the number of features detected having an isotope pattern in 17 plasma samples. We counted the number of masses having isotopic pattern in an artifact-free region between 350–700 *m/z* and 16–25 minutes (Fig. S2 in the Supporting Information). In the following paragraphs, we will show the results for a representative control sample comparing GridMass, Centroid, and Recursive Threshold from MZmine 2 and centWave from XCMS. Although Centroid and centWave were not designed for continuous mode and therefore their results may not be optimal, we used here those results as a comparative reference. Also, we will show the results of the analysis of the 17 plasma samples using GridMass.

Mass detection in a representative sample per method

The overall results of GridMass, Centroid, Recursive threshold, and centWave in a representative sample are shown in Fig. 3 and detailed in Table S2 (Supporting Information). In MZmine 2, GridMass was superior in the number of isotope-confirmed masses detected followed by Centroid then Recursive Threshold. GridMass generated at least 25% more masses than other algorithms in MZmine 2. The centWave algorithm was more sensitive to the height parameter and detected more masses than GridMass. However, a careful analysis shown in Fig. 3 (D) and Fig. S3 (Supporting Information) suggests that centWave could be generating a high number of false positives, perhaps because of the algorithm confusions in continuous mode data. Contrary, the Centroid method that was also designed for centroid data does not show such behavior. Comparing the detected masses across the three best methods [Fig. 3(B)], a high proportion of the masses were detected by at least two methods. The peak area of masses

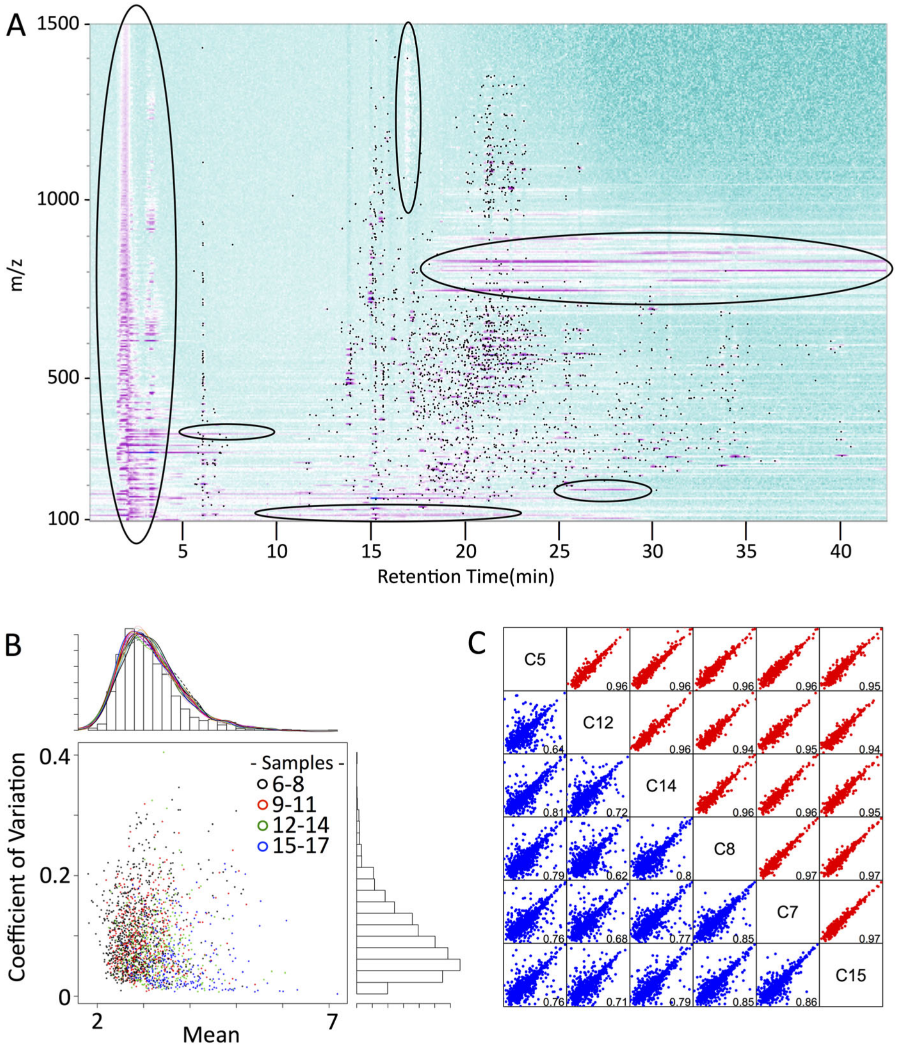

Figure 4. Masses detected by GridMass in 17 plasma samples. Panel A shows the masses detected within the 2D chromatogram. Ellipses denote examples of artifacts successfully ignored by the algorithm. Panel B shows the mean and coefficient of variation of masses detected along samples and their distribution in histograms. Lines over the top histogram represent the signal distribution of individual samples. Dot colors represent the number of samples where the mass was detected. Panel C shows scatter plots of six representative samples using all masses (lower triangle in blue) or using 50% less variable masses (upper triangle in red) of those masses whose standard deviation is less than 0.237 (to highlight between-sample correlation).

detected by any two methods is highly correlated across methods [Fig. 3(C)]. These results suggest that GridMass mass detection is also reliable in continuous mode data.

We also tested the effect of smoothing the chromatographic signal before the GridMass mass detection. Table S2 (Supporting Information) show that smoothing increases the detection of the number of isotopic masses for low values of the minimum height parameter (*height* = 20 or less) suggesting that smoothing may improve mass detection.

In addition, Table S3 (Supporting Information) provides an estimation of the overall feature detection across intensity thresholds. This table shows that, as expected, the percentage of masses included in isotope patterns decreases with the intensity threshold even when the total number of detected masses increases.

Runtime in a representative sample

Overall, the runtime varied depending on the algorithm and the parameters (Table S2, Supporting Information). However, GridMass was always faster than other algorithms. Moreover, GridMass runtime was always within the order of seconds. The running time of the centWave algorithm and the Centroid algorithm in MZmine 2 was highly variable, from less than a minute to almost 2 h showing increases for lower values of the height parameter. We noted that in the Centroid and Recursive Threshold algorithms in MZmine 2, the major time consuming task was the chromatographic builder step rather than the peak detection algorithm itself; therefore, the estimated time should be similar for changes on the detection algorithm in MZmine 2. Interestingly, the runtime of GridMass algorithm was almost insensitive to the height parameter. This is because the number of probes within the grid does not change. In addition, the time needed for smoothing (around 15 s) can be negligible.

Mass detection in the set of 17 plasma donors

When the best parameters (minimum height = 10, smoothing time = 0.05 min, smoothing *m/z* = 0.05, and ignore times = 0–5) were applied to the entire set (17 samples), we obtained 37 597 features after de-isotoping using GridMass. This represents an average of 2212 features per sample (*SD* = 158). After aligning, 55% of the features (n = 20 662) were detected in at least six of the 17 samples (Fig. S4 in the Supporting Information). For further analysis, we used only aligned features detected in at least six samples resulting in 2006 distinct features. From this, the average number of features per sample was 1215 (*SD* = 76). An overview of the final detected features is shown in Fig. 4. This figure shows that most of the detected features correspond to a putative 'spot'. Note that large artifacts are not included suggesting that artifact filtering was highly specific. To compare the intensity of features detected across samples, we used the Log_{10} of the area of the 'peak' estimated from MZmine 2. The Fig. 4(B) shows that the distribution of intensities was rather similar across samples. A correlation analysis shows highly correlated measurements [Fig. 4(C)]. An unsupervised cluster analysis shows that, apart from the sample C12, no major differences are observed.

These results suggest that GridMass yields consistent and correlated estimations across different samples.

Experiment 3: metabolic profiling of habanero pepper samples

In order to evaluate the performance of GridMass with more complex settings, we conducted a controlled experiment in plants. It is known that in general, plants contain orders of magnitude more metabolites than mammals.[11,12] Therefore, this experiment is expected to represent a challenging scenario for a peak detection tool. A crucial operation parameter, particularly in samples with a high peak count, is the ability to detect the most features while

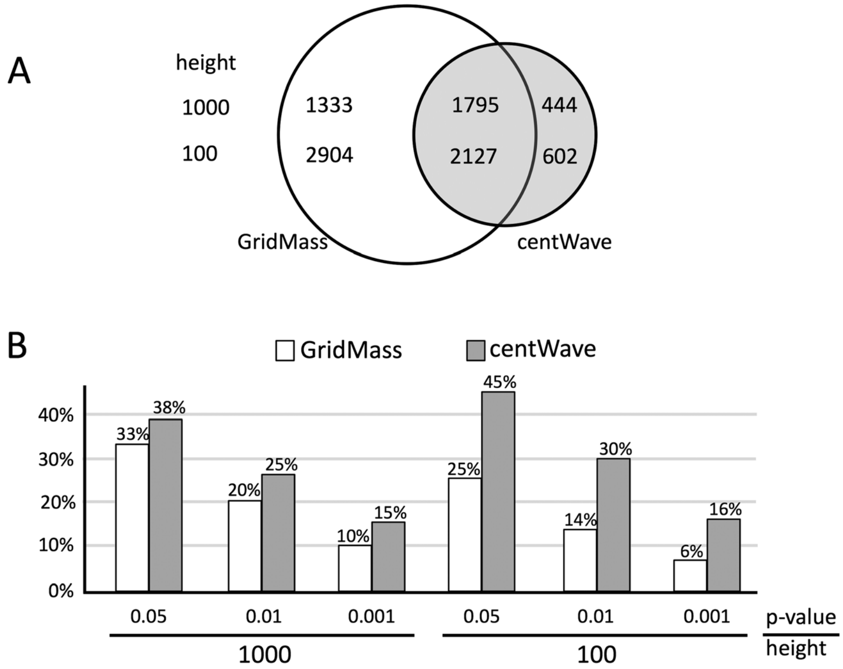

Figure 5. Comparison of features detected by GridMass and CentWave from the Habanero samples in Experiment 3. Panel A shows the number of features detected after de-isotoping by the two algorithms using two height thresholds. Panel B shows a comparison of the percentage of false positives for the two height thresholds at three p-values criteria that determines false calls.

avoiding irreproducible 'artifacts' that could yield false positives downstream in the data analysis process. To assess the robustness of the algorithms, we analyzed seven biological replicates of Habanero fruits comparing the differences between three technical replicates injected non-consecutively. Differences between technical replicates should represent false positives. Results are compared using GridMass and CentWave. The minimum peak heights used were 100 and 1000 for both algorithms. The chromatograms were recorded in centroid mode.

Overview of detected features

An overview of the whole chromatogram is shown in Fig. S5 (Supporting Information). As shown in Fig. 5, GridMass detected more features than centWave after de-isotoping independently of the value of the minimum height threshold. GridMass detected 3218 and 5031 de-isotoped features for minimum height of 1000 and 100, respectively, whereas centWave detected 2239 and 2729. After comparisons of the same detected features on both pipelines, GridMass detected between 78% and 80% of the features detected by centWave; however, centWave detected only between 42% and 57% of the features detected by GridMass. In these experiments, GridMass was faster than centWave.

False positives

We evaluated the false positive rate using, rANOVA, a statistical approach tailored for repeated measurements. The results shown in Fig. 5 clearly demonstrate that GridMass delivered less false positives than centWave independently of the minimum height threshold and the *p*-value cutoff. The difference in false positive rates between centWave and GridMass ranges from 5% to 20%. We observed that these differences were higher for height = 100.

These results show that GridMass is competitive and useful and can achieve better sensitivities and specificities than centWave.

Conclusion

GridMass is an efficient, fast, and competitive 2D feature detection algorithm for LC/MS analysis implemented in Java for the MZmine 2 software. Our results suggest that GridMass is a useful tool for HPLC/MS analysis within the MZmine 2 framework. We observed in three experiments using molecule standards, human plasma, and habanero fruits that GridMass can achieve comparable or higher sensitivities and specificities than other algorithms. We conclude that GridMass is a valuable tool for feature detection in metabolomic studies, which can be easily used by non-expert users. In the Supporting Information, we showed that GridMass can also be useful for proteomics. GridMass is available at http://bioinformatica.mty.itesm.mx/GridMass, in the MASSyPup distribution,[6] and in the forthcoming MZmine 2 versions.

Acknowledgements

We thank Gabriela Gutierrez Orozco for her help in obtaining plasma samples, Tomas Pluskal for MZmine assistance, and Dr. Sergio Lozano for writing corrections.

We thank the Tecnológico de Monterrey (ITESM) for funding I.L. Y.G., C.R.L., R.U.L. including the Cátedra de Bioinformática, the Cátedra de Micronutrientes, and Cátedra Empresarial de Tecnologías de Agricultura Intensiva. We also thank the Instituto Mexicano del Seguro Social (IMSS) for allowing the obtention of samples, the Consejo Nacional de Ciencia y Tecnología (CONACyT) for grant 140601, and scholarships for I.L.Y.G, C.R.L., and R.U.L.

References

[1] R. C. De Vos, S. Moco, A. Lommen, J. J. Keurentjes, R. J. Bino, R. D. Hall. Untargeted large-scale plant metabolomics using liquid chromatography coupled to mass spectrometry. *Nat. Protoc.* **2007**, *2*, 778.

[2] T. Pluskal, S. Castillo, A. Villar-Briones, M. Oresic. MZmine 2: modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**, *11*, 395.

[3] R. Tautenhahn, C. Bottcher, S. Neumann. Highly sensitive feature detection for high resolution LC/MS. *BMC Bioinformatics* **2008**, *9*, 504.

[4] J. Cox, M. Mann. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **2008**, *26*, 1367.

[5] M. Sturm, A. Bertsch, C. Gröpl, A. Hildebrandt, R. Hussong, E. Lange, N. Pfeifer, O. Schulz-Trieglaff, A. Zerck, K. Reinert, O. Kohlbacher. OpenMS–An open-source software framework for mass spectrometry. *BMC Bioinformatics* **2008**, *9*, 163.

[6] R. Winkler. MASSyPup–an 'out of the box' solution for the analysis of mass spectrometry data. *J. Mass Spectrom.* **2014**, *49*, 37.

[7] W. E. Haskins, K. Petritis, J. Zhang. MRCQuant-an accurate LC-MS relative isotopic quantification algorithm on TOF instruments. *BMC Bioinformatics* **2011**, *12*, 74.

[8] E. Kenar, H. Franken, S. Forcisi, K. Wormann, H. U. Haring, R. Lehmann, P. Schmitt-Kopplin, A. Zell, O. Kohlbacher. Automated label–free quantification of metabolites from liquid chromatography–mass spectrometry data. *Mol. Cell. Proteomics* **2014**, *13*, 348.

[9] D. Hoagland, D. Arnon. The Water-Culture Method for Growing Plants Without Soil. D. R. Hoagland (Ed.). Free Download & Streaming: Internet Archive, C347, College of Agriculture, University of California: Berkeley, California, **1950**. 1884–1949.

[10] M. A. Lawrence, ez: Easy analysis and visualization of factorial experiments. R package version *4.2-2*, **2013**.

[11] D. B. Kell. Systems biology, metabolic modelling and metabolomics in drug discovery and development. *Drug Discov. Today* **2006**, *11*, 1085.

[12] E. Pichersky, D. R. Gang. Genetics and biochemistry of secondary metabolites in plants: an evolutionary perspective. *Trends Plant Sci.* **2000**, *5*, 439.

Supporting information

Additional supporting information may be found in the online version of this article at the publisher's web site.