Analytica Chimica Acta 1171 (2021) 338669

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

rMSIannotation: A peak annotation tool for mass spectrometry imaging based on the analysis of isotopic intensity ratios

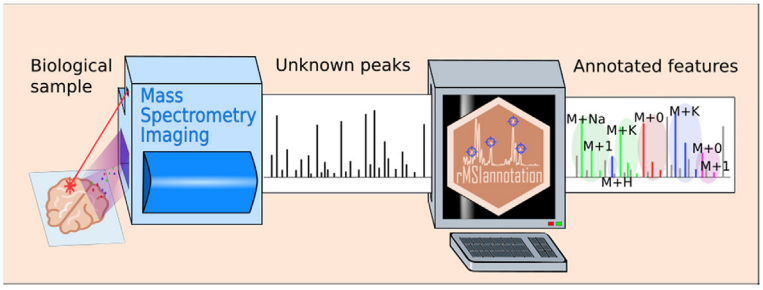
Lluc Sementé [a], Gerard Baquer [a], María García-Altares [a, b, *], Xavier Correig-Blanchar [a, b, c], Pere Ràfols [a, b, c]

[a] University Rovira I Virgili, Department of Electronic Engineering, Tarragona, Spain
[b] Spanish Biomedical Research Centre in Diabetes and Associated Metabolic Disorders (CIBERDEM), 28029, Madrid, Spain
[c] Institut D'Investigació Sanitària Pere Virgili, Tarragona, Spain

HIGHLIGHTS

- MSI Peak annotation tool without supporting libraries.
- Easy MSI data reduction to mono-isotopic ions.
- Discovery of adduct ions using isotopic pattern information.

GRAPHICAL ABSTRACT



ARTICLE INFO

Article history:
Received 22 February 2021
Received in revised form
12 May 2021
Accepted 20 May 2021
Available online 23 May 2021

ABSTRACT

Mass spectrometry imaging (MSI) consist of spatially located spectra with thousands of peaks. Only a fraction of these peaks corresponds to unique monoisotopic peaks, as mass spectra include isotopes, adducts and fragments of compounds. Current peak annotation solutions depend on matching MS features to compounds libraries. We present rMSIannotation, a peak annotation algorithm to annotate carbon isotopes and adducts in metabolomics and lipidomics imaging mass spectrometry datasets without using supporting libraries. rMSIannotation measures and evaluates the intensity ratio between carbon isotopic peaks and models their distribution across the *m/z* axis of the compounds in the Human Metabolome Database. Monoisotopic peak selection is based on the isotopic likelihood score (ILS) made of three components: image morphology correlation, validation of isotopic intensity ratios, and peak centroid mass deviation. rMSIannotation proposes pairs of peaks that can be adducts based on three scores: isotopic pattern coherence, image correlation and mass error. We validated rMSIannotation with three MALDI-MSI datasets which were manually annotated by experts, and compared the annotations obtained with rMSIannotation and with the METASPACE annotation platform. rMSIannotation replicated more than 90% of the manual annotation reported in FT-ICR datasets and expanded the list of annotated compounds with additional monoisotopic peaks and neutral masses. Finally, we evaluated isotopic peak annotation as a data reduction method for MSI by comparing the results of PCA and *k-means* segmentation before and after removing non-monoisotopic peaks. The results show that monoisotopic peaks retain most of the biologic variance in the dataset.

* Corresponding author. University Rovira i Virgili, Department of Electronic Engineering, Tarragona, Spain.
E-mail address: maria.garcia-altares@urv.cat (M. García-Altares).

https://doi.org/10.1016/j.aca.2021.338669

1. Introduction

Mass spectrometry imaging (MSI) is a technique that can spatially resolve the chemical composition of a variety of biosamples, including animal and plant tissues, to reveal their biological mechanisms [1–3]. An MSI dataset consists of a collection of mass spectra localized in the pixels of an image. Raw mass spectra need to be processed to reduce the variance introduced during acquisition (electronic noise, mass drifts, intensity fluctuations, etc.) [4]. The information in a processed dataset consists of spatially resolved discrete *m/z* features, which undergo data analysis steps such as multivariate statistics and compound identification to obtain biological knowledge [5–7].

Mass spectrometry dataset contains redundant information, since a single chemical compound generates multiple peaks, which can be attributed to isotopes, adducts, fragments, and different ionization states. Therefore, the redundant variables in the dataset tend to enlarge the data size and hinder statistical analysis [8]. Reducing this redundancy to obtain statistically relevant variables is crucial to unveiling biological knowledge [9–11].

In this study, we define peak annotation as the process of automatically grouping all peaks related to the same molecule, and the ion species to which they correspond [11–13]. This involves labeling carbon monoisotopic (M+0) and isotopic ions (M+1, M+2, etc.), adducts of the same compounds ($[M+Na]^+$ $[M+K]^+$, etc.), and, when possible, assigning putative molecular classes with the Kendrick mass defect [14,15]. Besides, a neutral monoisotopic mass can be determined if two or more adducts can be annotated for a given compound. This allows the assignment of molecular formulas with higher confidence. Peak annotation is an essential step prior to peak identification, which consists in searching the annotated peaks in libraries of chemical compounds to assign them a putative chemical formula and name using MS data and confirming each assignment through MS^n data and orthogonal techniques [16].

Moreover, peak annotation algorithms are reliable variable selection approaches and greatly facilitate the identification process. Annotation ideally allows unifying all peaks coming from the same compound, reducing the number of statistical variables to only one independent variable per compound.

Peak annotation algorithms are more established in LC-MS-based experiments than in MSI. Although, LC-MS and MSI have different data structure and content, some peak annotation strategies in LC-MS can be adapted to MSI datasets. Notable examples are:

(1) R package CAMERA [11] annotates carbon isotopes, adducts, and fragments in a peak list by first grouping peaks by peak shape correlation, retention time similarity and correlation across samples, and then by checking M+1/M+0 isotopic ratios, and adduct distances. Ratios between M+0 and M+1 isotopes are computationally pre-established.
(2) R package CliqueMS [17] annotates adducts using the similarity between coelution profiles and a similarity network based on the natural frequency of adduct formation observed in real samples.
(3) R package Astream [18] annotates isotopes, fragments, and adducts by using intensity correlations across samples, retention time differences, and expected *m/z* differences.

In MSI there is no chromatographic separation before ionization and ions frequently overlap, even with high resolving power spectrometers (>20,000). Since MSI is an imaging technique, spatial correlation methods can be used to increase peak annotation confidence. To our knowledge, only two annotation tools have been developed specifically for MSI applications:

(1) R package MassPix [19] annotates M+0, M+1 and M+2 isotopes by searching for intensity ratios between peaks below user defined ratios. After deisotoping, it searches for the *m/z* of M+0 peaks in a self-developed library of lipids to tentatively annotate and identify them. MassPix does not consider spatial information or colocalization among isotopic ion images.
(2) METASPACE annotation platform [20] is an online annotation tool in which users upload their MSI datasets to be annotated. Its annotation workflow consists of generating isotopic patterns from metabolites databases and matching them with the experimental MSI data using three different metrics: spatial chaos measure, spatial isotope measure and spectral isotope measure. Matches with an overall score higher than a threshold are then given a false discovery rate score based on a target-decoy approach [21]. The results of this workflow are pairs of matching adducts and formulae, which lead to tentative *m/z* identifications. On the downside, it is important to notice that METASPACE requires to uploads datasets with a high mass accuracy (<3 ppm) and a resolving power over 70k (*m/z* 200) for reliable results. In addition, METASPACE may be impractical for large experiments since datasets must be uploaded through the internet. Finally, despite having METASPACE 's source code available, it still suffers from the black box effect where users are restricted to visualize the annotation results and are not able to finely control/adapt the annotation tool themselves.

Both MassPix and METASPACE use generated isotopic patterns from libraries of metabolites, which restrict the annotation to compounds already reported in the libraries. To overcome this limitation, we propose rMSIannotation, a new annotation tool based on the analysis of isotopic patterns optimized for compounds below 1200 Da, included in the MSI data processing R package rMSIproc [22]. rMSIannotation takes advantage of the high number of pixels in an MSI dataset to annotate carbon-based isotopes with single and multiple charges using three scores: (1) image morphology, which considers the colocalization among related *m/z* ion images, (2) isotopic pattern profile, which asserts the plausibility of isotopic ratios given an *m/z* ratio and (3) centroid mass deviation, which evaluates the theoretical distance of carbon isotopic patterns. Additionally, monoisotopic ions found by the algorithm are compared with theoretical mass distances of adducts to generate tentative neutral masses. The algorithm has been tested and validated using *in silico* datasets, experimental datasets with manual identifications and by comparing the annotations produced by rMSIannotation with the results provided by METASPACE. Users can freely access and/or contribute to rMSIproc at <https://github.com/prafols/rMSIproc>.

2. Materials and methods

2.1. Imaging datasets

Three published datasets were used to test the algorithm: (1) a MALDI-TOF dataset consisting of bovine ovarian follicles [23], (2) a MALDI-FT-ICR dataset consisting of a bloom-forming alga during infection [24] and (3) a MALDI-FT-ICR dataset consisting of coronal 12 μm-thick brain sections of adult wild-type C57 mice [20].

2.1.1. MALDI-TOF dataset

The MALDI-TOF dataset consists of a collection of bovine ovarian follicles [23]. The dataset was kindly provided by the authors. Details of sample preparation and data acquisition can be found in the original paper. The authors identified 43 metabolites in the MSI dataset by first, analyzing lipid extracts from the follicular cells with

high-resolution LC-MS and direct infusion MS/MS structural analyses and second, searching the identifications in the MSI dataset. The raw data was exported to imzML format using Bruker FlexImaging software and the dataset was then processed using the rMSIproc processing workflow [22]. The processing pipeline consisted of: (1) smoothing by Savitzky-Golay using a kernel size of 7, (2) spectra alignment with two iterations, a 400 ppm max shift, an oversampling of 2 and references for low, mid and high of 0, 0.5 and 0.8, (3) mass calibration using previously identified peaks (*m/z* 524.372, *m/z* 760.586 and *m/z* 824.557) and (4) peak-picking with an SNR threshold set to 5, a peak detector window of 12, a peak oversampling of 10, a binning tolerance of 5 scans and a binning filter of 0.05. The result was a peak matrix with a total of 235 peaks and 15293 pixels within the *m/z* range between 100 and 1200.

2.1.2. MALDI-FT-ICR dataset 1

The MALDI-FT-ICR dataset 1 consists of a bloom-forming alga (Emiliana huxleyi) during infection with a virus [24]. The dataset was available from MetaboLights [25] stored in the study with reference MTBLS769. Details of sample preparation and data acquisition can be found in the original paper, in which the authors identified 37 metabolites using LC-MS and LC-MS/MS in lipidomic experiments performed in liquid cultures. The raw data was exported to imzML format using Bruker FlexImaging and the dataset was then processed using rMSIproc. The processing pipeline consisted of: (1) smoothing by Savitzky-Golay using a kernel size of 7, (2) a spectra alignment with two iterations, a 300 ppms max shift, an oversampling of 2 and references for low, mid and high of 0, 0.5 and 1, (3) mass calibration using four previously identified peaks (*m/z* 689.5024, *m/z* 749.5153, *m/z* 802.5469, *m/z* 826.6199 and *m/z* 902.5782) to facilitate the comparison of the results and (4) peak-picking with an SNR threshold set to 20, a detector window of 10, an oversampling of 10, a binning tolerance of 6 scans and a binning filter of 0.05. The result was a peak matrix with a total of 4047 peaks and 10517 pixels within the *m/z* range of 100−1200.

2.1.3. MALDI-FT-ICR dataset 2

The MALDI-FT-ICR dataset 2 consists of four coronal 12 μm-thick brain sections of an adult wild-type C57 mice [20]. The dataset was available from MetaboLights [25] stored in the study with reference number MTBLS313. Details of sample preparation and data acquisition can be found in the original paper. In the original work, the dataset consisted of ten sections of two different animals. In this work, we used four sections out of five from the first animal as the data for one section was missing. The authors annotated 35 molecules for the first animal using the METASPACE platform and validated 16 representative annotations with LC-MS/MS.

The data was obtained from individual imzML files in processed mode containing the peaks list of each section, which was transformed to rMSIproc's peak matrix format using a mass binning of 10 ppms and a bin filter of 1%. After that, the four peak matrices were combined in a single dataset using rMSIproc's processing pipeline. The resulting peak matrix contained 1011 peaks and 53241 pixels within a mass range from *m/z* 100 to *m/z* 1180.

2.2. Description of the algorithm

The algorithm consists of two modules: isotope annotation and adduct annotation. The isotope annotation module detects pairs of isotope candidates and computes the isotopic detection metrics for all the peaks in the dataset. The adduct annotation module use the information generated by the isotope annotation module and proposes pairs of peaks that could be adducts of the same compound. Lastly, all the annotations generated are organized in three groups: two groups for the adduct module, differentiated by the amount of information gathered during the annotation; and one for the isotope module containing information on the monoisotopic ions. Fig. 1 shows a flow diagram of the algorithm.

2.2.1. Input data format

Raw spectra undergo rMSIproc's processing workflow [22], which consists of spectral smoothing, spectral alignment, mass recalibration, peak picking, and peak binning of all the pixels in the image. The result of this workflow is a peak matrix, in which pixels of the image are arranged in rows, *m/z* features are arranged in columns and the *m/z* axis is shared between all pixels.

The annotation algorithm uses the rMSIproc peak matrix format as input. Alternatively, rMSIproc can create a peak matrix from an imzML file already centroided by third-party software. However, it is recommended to use raw data in profile mode to take full advantage of the complete rMSIproc processing workflow.

2.2.2. Isotope annotation

First, all *m/z* features in the peak matrix are assumed to be M+0 ions and, for all of them, a list of possible M+1 candidates is generated looking for peaks at a mass distance of 1.00336 Da within a user-defined windows (depending on the spectral resolution of the MS analyzer), expressed in number of raw spectra data points. We prefer to specify this mass distance in data points instead of ppm since it provides a more constant metric thought all the mass range. Alternatively, if spectral data is not available in profile mode, the mass tolerance can be specified in ppm. The mass distance is divided by the charge number, if isotopes of ions with multiple charges are being searched for.

Next, the *m/z* features with one M+1 candidate or more are evaluated pairwise with the isotopic likelihood score (ILS) which was developed in-house and consists of the combination of three different scores: 1) the image morphology score, 2) the isotopic pattern profile score and 3) the centroid mass deviation score. Before computation, the pixels with zero value are removed pairwise from both *m/z* features to increase the discriminant power of the score.

1. The image morphology score considers that *m/z* features belonging to the same isotopic pattern are colocalized. We estimate colocalization by least squares regression between the intensities of M+0 and the M+1 candidate across all the pixels using the coefficient of determination (R^2). Ions are colocalized if the coefficient is close to 1.
2. The isotopic pattern profile score examines the relationship between the experimental and the theoretical M+1/M+0 intensity ratios. The experimental intensity ratio is defined as the slope of a linear model produced by least squares regression between the M+0 and the M+1 candidate intensities. The theoretical intensity ratio is calculated inputting the *m/z* of the monoisotopic candidate to a self-developed carbon isotopic ratio model (CIR model). The carbon isotopic ratio model contains the distribution of carbon isotopes intensities ratios across the *m/z* axis up to *m/z* 1200 and delivers the most probable intensity ratio for a given peak mass (see section 1 of supplementary information). Lastly, the experimental and theoretical intensity ratios are subtracted and fitted in a Gaussian score function which preserves the expected variability of the carbon isotopic ratio model. The score gets close to one as the measured intensity ratio of a pair of peaks is more likely to result from an actual isotopic profile.
3. The centroid mass deviation score compares the experimental mass distance between M+0 and its M+1 with the theoretical mass distance between carbon isotopes (considering the

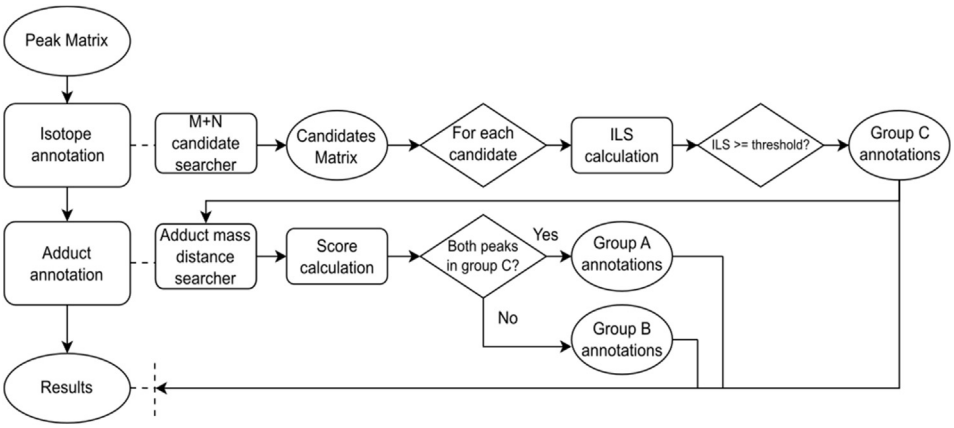


Fig. 1. Flow diagram of the peak annotation algorithm rMSIannotation. Rounded objects refer to data structures and rectangles to algorithmic processes.

charge). The user defines the error tolerance for the mass deviation, which can be introduced in ppms or number of data points. The score gets close to one as the error tolerance reduces.

The three scores are multiplied to calculate the ILS. The pairs of *m/z* features with an ILS greater than the user-defined threshold constitute a monoisotopic/isotopic peak pair. Once all the true M+0 *m/z* features have been found, the full procedure is repeated to evaluate the M + N candidates for all the M+0 *m/z* features until no more candidates are found or N has reached the maximum number of iterations. The number of isotopes (N) to search for is a user-defined parameter.

2.2.3. Adduct annotation

The algorithm searches for pairs of ions (discarding the features annotated as isotopes) whose mass difference fits with a candidate adduct ($[M+H]^+$, $[M+Na]^+$, $[M+K]^+$, user-defined adducts, and neutral losses) within a mass tolerance in ppms to generate putative neutral masses. For each pair of adduct ions, the algorithm calculates three scores to guide the user to select the more probable adduct pairs. The scores are:

1. Isotopic pattern coherence. When two monoisotopic ions are adducts of the same compound, their M+1/M+0 intensity ratio should be the same (unless the ion forming the adduct contains carbon, which would slightly modify the isotopic pattern). This is calculated as the standard error of the mean M+1/M+0 intensity ratios of both monoisotopic ions. Small standard error of the mean indicates good isotopic pattern coherence.
2. Correlation between the two ions intensities using Pearson's R. We assume that adducts of the same compound exhibit some degree of colocalization. It is expected to obtain less degree of colocalization between adducts peaks than between isotopes peaks due to salts concentrations variations related to tissue morphology. Nevertheless, the ion images between adducts of the same compound should be still similar and very rarely show complementary spatial distributions.
3. Mass error between the M+0 peaks and their putative neutral mass. The neutral mass is calculated by subtracting the molecular mass of each adduct ion and averaging the resulting neutral masses. Small mass errors indicate a precise putative neutral mass assignation.

The algorithm allows each *m/z* feature to be part of different adduct pairs (e.g., an $[M+Na]^+$ ion can be paired with an $[M+H]^+$ ion and with an $[M+K]^+$ ion) and even labeled as different adducts in different pairs (e.g., an ion can be labeled as $[M+Na]^+$ in one pair and as $[M+K]^+$ in a different pair). The calculated scores of each annotation are stored along with each adduct pair, which enables the user post-evaluation of all possible adducts pairs to select the most feasible annotations. The user is the responsible to choose/validate the more feasible annotations provided by the algorithm.

Finally, the adduct annotation module generates a list with the neutral masses and its annotation scores, facilitating the search in compound libraries for tentative identification.

2.2.4. Feature annotation groups and output information

The annotations are divided into three groups (A, B and C) depending on the information available to reliably annotate each *m/z* feature.

Group 'A' contains neutral masses from pairs of M+0 ions cataloged as adducts, where at least one isotope is identified for every M+0 ion. The three scores described for adducts can be computed for all these pairs.

Group 'B' contains neutral masses from pairs of ions in which, one ion is an M+0, but not the other. The isotopic information is not available for the second ion since the algorithm failed to assign the corresponding M+1 peak. Therefore, isotopic pattern coherence cannot be computed in this annotation group.

Group 'C' contains the *m/z* ratios of all M+0 annotated ions. Ions only reported in group C are, therefore, annotated as M+0, but their adduct identity is unknown. This group consists of a summary of the isotope annotation module, in which ILS is the key quality parameter.

The output of rMSIannotation consists of the annotations in groups A, B and C (which can be exported as CSV files); the computations of the ILS for all candidates during isotope annotation, and two vectors of the monoisotopic and isotopic ions. The vectors of monoisotopic and isotopic ions can be used to filter the peak matrix to remove the isotopic peaks, or to work with only the monoisotopic ions found.

3. Results

First, we tested the performance of rMSIannotation using two *in silico* MSI datasets. The datasets were developed simulating TOF and FT-ICR mass analyzers experiment in which we know a priori the identity of all the *m/z* features. Section 2 of supplementary information contains the detailed procedure. Then, we used different ILS thresholds with the *in silico* datasets to test the performance of rMSIannotation's criteria and to obtain optimal ILS thresholds. The optimal ratios found were 0.55−0.7 for TOF datasets and 0.7 to 0.8

for FT-ICR datasets. Next, we compared the number of coinciding annotations produced by sweeping the ILS threshold in a range of 0.2−0.9 for the MALDI-TOF dataset and MALDI-FT-ICR dataset 1. This allowed us to determine whether the optimal ILS thresholds obtained with the *in silico* dataset were applicable to experimental data. The results show that the number of annotations provided by rMSIannotation coinciding with the manual annotations decreases slowly as we increase the ILS threshold until it reaches the optimal thresholds (Fig. S5). After this point, the number of coinciding annotations drastically decreases. This suggests that the optimal ILS thresholds obtained *in silico* are applicable experimental data and can be setup as default parameter values. Refer to Section 3 of supplementary information for the complete study.

Second, we annotated using rMSIannotation three experimental datasets acquired with TOF and FT-ICR mass analyzers from papers that reported manually identified compounds. We compared the reported annotations with the ones generated by rMSIannotation. Later, we compared the annotations of rMSIannotation with the results obtained using METASPACE annotation platform on the FT-ICR datasets.

Finally, we evaluated the effects of retaining only M+0 ions during the post-processing of MSI datasets, using principal component analysis (PCA) and *k-means* clustering.

3.1. MALDI-TOF annotation results

The MALDI-TOF dataset consists of a collection of bovine ovarian follicle tissues in which the authors identified 43 metabolites (see Fig. 2). After the raw data had been processed, the peak matrix was fed to the annotation algorithm. The parameters used were: isotope search up to M+3, isotope mass tolerance in data points mode and up to 4 data points (~100 ppms at *m/z* 800 for this dataset), ILS threshold set to 0.6 and default $[M+K]^+$, $[M+H]^+$ and $[M+Na]^+$ adducts searched for within a window of 30 ppm.

With these parameters, rMSIannotation generated 16 putative neutral masses in group A and 22 in group B, and found a total of 42 monoisotopic ions in group C. All the annotations of each group are presented in Supplementary Table S1, S2 and S3.

First, we compared the adduct ions found in groups A and B with the adduct ions from the original publication. This was done by searching in groups A and B for the exact masses of the compounds identified. Then, we searched for monoisotopic ions without adduct annotation in group C. Table 1 shows the monoisotopic ions found by rMSIannotation that coincide with those identified in the original work. The ions in group C that also appear in groups A or B (annotated as monoisotopic ions) display its ILS value.

We annotated as monoisotopic 23 of the ions in the list of 43 provided by the authors in the original study (see Fig. 2). There are three causes explaining why the other 20 ions provided by the authors were not annotated as monoisotopic ions by rMSIannotation: (1) the peak picking algorithm detected only the M+0 ion due to low intensity of the subsequent isotopes; (2) all the ions of the compound have an intensity group below the S/N ratio, and (3) overlapping isotopic patterns of isobaric species which could not be properly resolved by the mass analyzer. Causes 1 and 2 are related to the presence of only one peak per compound in the MSI dataset as the provided identifications were obtained using LC-MS and direct infusion MS/MS. In addition, we further analyzed the case of overlapping with *in silico* overlapping isotopic patterns with different resolving power to determine how it affects rMSIannotation (section 4 of supplementary information). The results show that, rMSIannotation is tolerant to some extend of peak overlapping and the resulting annotation depend on the two overlapped compounds abundance ratios and on the spectral resolving power. As expected, a higher resolving power increases the annotation performance, but even when lowering the resolving power the algorithm still provides reliable results by annotating peak in the isotopic pattern (M+1, M+2, …) only when isotopic ratio criteria is met. Therefore, monoisotopic peaks (M+0) highly overlapped with the M+1 peak of another molecule will not be annotated as part of an isotopic pattern of the former molecule. Supplementary Table S4 shows which category applies to the non-annotated ions and Supplementary Figures S9, S10 and S11 show examples of each group defined above, respectively. It is worth mentioning that some of the non-annotated ions could have been annotated by reducing the SNR in the preprocessing steps of the datasets although uninformative noisy peaks may be introduced hampering the subsequent data analysis.

Lastly, we used the Human metabolome database [26] and Lipid maps [27] to putatively identify the ions annotated by rMSIannotation that had not been identified in the original paper. We identified 1 neutral mass with a mass error below 30 ppms that belonged to the CHCA molecule used as matrix (we found 9 common adduct ions by hand in group C), and 4 more monoisotopic masses, resulting in 13 new monoisotopic ions identified. Supplementary Table S5 shows the putative name and molecular formula for the 4 monoisotopic masses in group C (CHCA related annotations are excluded).

3.2. MALDI-FT-ICR annotation results

The MALDI-FT-ICR dataset 1 consists of a bloom-forming alga (*Emiliana huxleyi*) analyzed during a viral infection [24]. The authors of the original paper identified 37 metabolites. The algorithm parameters used were: isotope search up to M+3, isotope mass tolerance in ppm mode up to 10 ppms, ILS threshold set to 0.7 and $[M+K]^+$, $[M+H]^+$ and $[M+Na]^+$ adducts searched up to a maximum of 5 ppm mass tolerance.

With these parameters, rMSIannotation generated 31 putative neutral masses in group A and 95 putative neutral masses in group B, and found a total of 187 monoisotopic ions in group C. All the annotations of each group are presented in Supplementary Table S6, S7 and S8.

Considering all the matching annotations, we found 28 ions on the list of 37 provided by the authors of the original study (Fig. 2) and we obtained 2 additional adducts for two of the compounds in the original work annotation list. Table 2 shows all the coinciding annotations. We observed that in this dataset several M+1 peak (and subsequent isotopes) have some pixels with null value due to the data reduction mode for FT-ICR raw data which automatically discarded low intensity signals. This produces a bias in the isotopic pattern profile score which can increase or decrease the real ILS score. To solve this problem, the algorithm is designed to discard pairwise pixels with null values to ensure proper linear modelization. Supplementary Fig. S12 shows the example of ion *m/z* 826.620 corresponding to compound DGCC 40:7, in which the ILS is 0.877 if null pixels are included and 0.984 if null pixels are discarded.

rMSIannotation was not able to annotate 9 of the manually identified compounds because of their low intensity. This means that the M+1 and subsequent isotopes were not present in the peak matrix or there were too many null pixels to be properly corrected by the algorithm. Supplementary Table S9 shows these compounds and Supplementary Fig. S13 shows the case of ions *m/z* 826.640 and *m/z* 812.622.

Various compound libraries were used to tentatively assign the new annotations generated by rMSIannotation not reported in the original work. Supplementary Table S10 shows the putative names and molecular formulae assigned to 19 monoisotopic masses, according to METLIN [28], Lipid Maps [27] and Dictionary of Natural Products [29]. It is worth mentioning that rMSIannotation helped

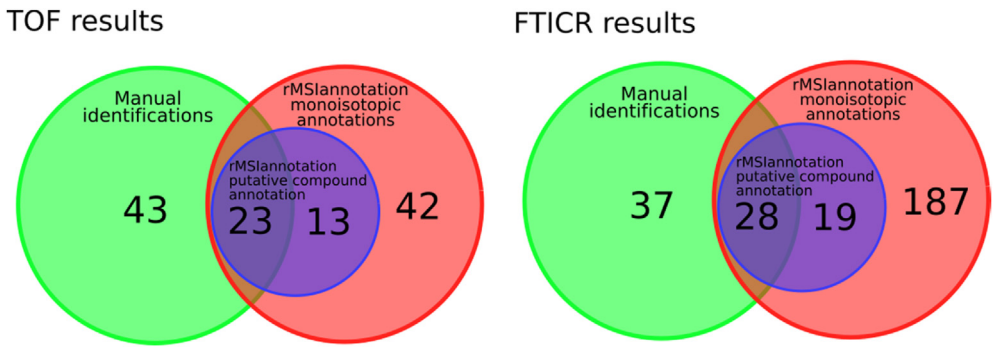


Fig. 2. Diagrams representing the number of identifications reported by the authors of the datasets, the number of M+0 annotations produced by rMSIannotation in group C and the number of coinciding and new putative compound annotations found using rMSIannotation.

Table 1
Coinciding annotations of MALDI-TOF dataset between rMSIannotation and author's manual identifications.

Name	Formula	Adduct	*m/z*	Mass error (ppm)	Annotation group	ILS
Phosphocholine	$C_5H_{15}NO_4P$	$[M]^+$	184.141	364.560	C	0.842
LPC 16a:0	$C_{24}H_{50}NO_7P$	$[M+H]^+$	496.367	54.866	A	0.848
LPC 16a:0	$C_{24}H_{50}NO_7P$	$[M+Na]^+$	518.348	50.717	B	–
LPC 18a:1	$C_{26}H_{52}NO_7P$	$[M+H]^+$	522.387	60.460	B	0.812
LPC 18a:0	$C_{26}H_{54}NO_7P$	$[M+H]^+$	524.372	1.781	C	0.746
LPC 16a:0	$C_{24}H_{50}NO_7P$	$[M+K]^+$	534.318	41.833	A	0.662
LPC 18a:1	$C_{26}H_{52}NO_7P$	$[M+Na]^+$	544.363	47.099	B	–
LPC 18a:1	$C_{26}H_{52}NO_7P$	$[M+K]^+$	560.338	47.653	B	–
SM(d18:1/C16:0)	$C_{39}H_{79}N_2O_6P$	$[M+H]^+$	703.576	1.632	B	–
SM(d18:1/C16:0)	$C_{39}H_{79}N_2O_6P$	$[M+Na]^+$	725.543	19.014	A	0.853
PC 32a:0	$C_{40}H_{80}NO_8P$	$[M+H]^+$	734.562	10.116	A	0.866
SM(d18:1/C16:0)	$C_{39}H_{79}N_2O_6P$	$[M+K]^+$	741.510	27.960	A	0.916
PC 32a:0	$C_{40}H_{80}NO_8P$	$[M+Na]^+$	756.526	33.542	A	0.672
PC 34a:2	$C_{42}H_{80}NO_8P$	$[M+H]^+$	758.548	28.253	B	0.705
PC 34a:1	$C_{42}H_{82}NO_8P$	$[M+H]^+$	760.574	14.569	A	0.829
PC 32a:0	$C_{40}H_{80}NO_8P$	$[M+K]^+$	772.508	22.411	A	0.641
PC 34a:2	$C_{42}H_{80}NO_8P$	$[M+Na]^+$	780.537	18.418	B	–
PC 34a:1	$C_{42}H_{82}NO_8P$	$[M+Na]^+$	782.539	35.814	A	0.877
PC 36a:2	$C_{44}H_{84}NO_8P$	$[M+H]^+$	786.585	19.999	A	0.692
PC 34a:2	$C_{42}H_{80}NO_8P$	$[M+K]^+$	796.522	4.159	B	–
PC 34a:1	$C_{42}H_{82}NO_8P$	$[M+K]^+$	798.517	30.009	A	0.866
PC 36a:2	$C_{44}H_{84}NO_8P$	$[M+Na]^+$	808.555	34.229	B	–
PC 36a:2	$C_{44}H_{84}NO_8P$	$[M+K]^+$	824.530	32.277	A	0.642

*A missing ILS value correspond to ions where the isotopic pattern could not be annotated and are exclusively in group B.

to find different adducts of common alkenones produced by *Emiliana huxleyi* [30,31].

Additionally, we submitted the MALDI-FT-ICR dataset 1 to METASPACE to compare its performance against rMSIannotation. Table 3 lists all the manually identified compounds by the authors of the datasets and shows which ions were annotated by rMSIannotation and/or METASPACE. In case of METASPACE, we show the results for FDR 10%, which are showcased as the default results in the online platform, and the results for FDR 20%. The libraries selected in METASPACE were the Human Metabolome Database, Lipid Maps and Chemical Entities of Biological Interest. For FDR 10%, taking as a reference the manually identified compounds, METASPACE found 12 coinciding monoisotopic ions, and for FDR 20% found 20, which is less than the 28 found by rMSIannotation.

To further compare the performance of rMSIannotation with the METASPACE annotation platform, we tested rMSIannotation with the MALDI-FT-ICR dataset 2, consisting of four coronal brain sections of two adult wild-type C57 mice, which were previously annotated by the authors of METASPACE, reporting 31 compounds. The parameters used with rMSIannotation for the MALDI-FT-ICR dataset 2 were: ILS threshold set to 0.7, isotope mass tolerance in ppm mode up to 5 ppms and $[M+K]^+$, $[M+H]^+$ and $[M+Na]^+$ adducts searched up to a maximum of 5 ppm mass tolerance. rMSIannotation was able to putatively identify all the 31 compounds annotated using METASPACE. Moreover, rMSIannotation found 202 monoisotopic ions in group C and a total of 263 neutral masses combining groups A and B. Table 4 show the lists of the 31 annotated ions, together with its ILS values. We obtained ILS values over 0.9 for every annotated compound indicating high confidence in the annotation and confirming the original METASPCAE results.

3.3. *Effect of reducing variables to monoisotopic ions in multivariate analysis*

In section 3.2 we have shown the ability of rMSIanotation to annotate monoisotopic ions and, thereby, to annotate the isotopes which carries redundant information. There are also many peaks that are not annotated that could correspond to overlapped peaks, matrix derived peaks, etc. To test if rMSIannotation annotates most of the relevant peaks in the datasets, we compared the results of Principal Component Analysis (PCA) and image segmentation on the complete dataset against the results using only the monoisotopic ions. The results of the multivariate analysis are similar in both cases, indicating that the set of monoisotopic peaks annotated

Table 2
Coinciding annotations of MALDI-FT-ICR dataset 1 between rMSIannotation and author's manual identifications.

Name	Formula	Adduct	*m/z*	Mass error (ppm)	Annotation group	ILS
Sulfonioglycerolipid 28:0	$C_{38}H_{72}O_8S$	$[M+H]^+$	689.502	0.710	C	0.959
Sulfonioglycerolipid 30:0	$C_{40}H_{76}O_8S$	$[M+H]^+$	717.534	0.382	C	0.935
DGCC 36:6	$C_{46}H_{77}NO_8$	$[M+H]^+$	772.573	0.653	C	0.996
PC 36:6	$C_{44}H_{76}NO_8P$	$[M+H]^+$	778.538	1.732	C	0.955
DGCC 37:6	$C_{47}H_{79}NO_8$	$[M+H]^+$	786.588	1.376	C	0.992
Sulfonioglycerolipid 36:6	$C_{46}H_{76}O_8S$	$[M+H]^+$	789.533	1.125	C	0.984
PDPT 36:6	$C_{44}H_{75}O_8PS$	$[M+H]^+$	795.500	1.167	C	0.972
TG 46:1	$C_{49}H_{92}O_6$	$[M+Na]^+$	799.679	1.309	C	0.943
PDPT 37:6	$C_{45}H_{77}O_8PS$	$[M+H]^+$	809.516	1.472	C	0.836
Sulfonioglycerolipid 38:6	$C_{48}H_{80}O_8S$	$[M+H]^+$	817.565	0.784	C	0.975
PDPT 38:6	$C_{46}H_{79}O_8PS$	$[M+H]^+$	823.530	1.524	C	0.941
DGCC 40:7	$C_{50}H_{83}NO_8$	$[M+H]^+$	826.620	0.632	C	0.984
TG 48:1	$C_{51}H_{96}O_6$	$[M+Na]^+$	827.710	1.148	C	0.919
PC 40:7	$C_{48}H_{82}NO_8P$	$[M+H]^+$	832.585	1.183	C	0.767
Sulfonioglycerolipid 40:7	$C_{50}H_{82}O_8S$	$[M+H]^+$	843.579	3.861	C	0.879
TG 50:6	$C_{53}H_{90}O_6$	$[M+Na]^+$	845.664	0.376	C	0.735
PDPT 40:7	$C_{48}H_{81}O_8PS$	$[M+H]^+$	849.547	0.663	C	0.971
TG 50:2	$C_{53}H_{98}O_6$	$[M+Na]^+$	853.726	1.619	C	0.805
PC 44:12	$C_{52}H_{80}NO_8P$	$[M+H]^+$	878.570	0.524	C	0.968
PDPT 42:9	$C_{50}H_{81}O_8PS$	$[M+H]^+$	895.530	1.899	C	0.856
TG 54:7	$C_{57}H_{96}O_6$	$[M+Na]^+$	899.710	1.531	B	0.966
BLL 44:12	$C_{54}H_{79}NO_{10}$	$[M+H]^+$	902.578	0.264	C	0.796
TG 56:7	$C_{59}H_{100}O_6$	$[M+H]^+$	905.759	0.295	B	–
TG 54:7	$C_{57}H_{96}O_6$	$[M+K]^+$	915.685	1.256	B	–
TG 56:7	$C_{59}H_{100}O_6$	$[M+Na]^+$	927.742	1.311	B	0.899
TG 58:16	$C_{61}H_{86}O_6$	$[M+Na]^+$	937.634	0.080	C	0.910
TG 58:12	$C_{61}H_{94}O_6$	$[M+Na]^+$	945.695	1.332	C	0.883
TG 58:11	$C_{61}H_{96}O_6$	$[M+Na]^+$	947.711	0.458	C	0.907
TG 58:10	$C_{61}H_{98}O_6$	$[M+Na]^+$	949.727	1.332	C	0.914
TG 58:9	$C_{61}H_{100}O_6$	$[M+Na]^+$	951.743	1.310	C	0.888

*A missing ILS value correspond to ions where the isotopic pattern could not be annotated and are exclusively in group B.

by rMSIannotation retains the relevant biological information.

We standardized the data and then we compared the PCA scores of the complete dataset with the PCA scores of a reduced version of the dataset containing only monoisotopic ions. This involves selecting 17.87% (42 out of 235) of the variables for the TOF dataset and 4.62% (187 out of 4047) of the variables for the FT-ICR dataset 1. We compared the images produced by the three principal components on the tissue in each case. We compared the spatial structures displayed in the principal components images of the complete and the reduced datasets by computing its similarity using Pearson's R correlation. For the TOF dataset the correlations were: R = 0.99 (PC1), R = 0.96 (PC2), and R = 0.90 (PC3); for the FT-ICR dataset 1 the correlations were: R = 0.91 (PC1), R = −0.87 (PC2), R = 0.90 (PC3). In both cases, the principal components exhibit a very similar distribution. Fig. 3 shows the images of the first three principal components encoded in RGB color space for each studied case. As it can be seen, the tissue morphology is preserved in the reduced dataset.

Next, we analyzed the importance of monoisotopic peaks in the loadings of the first two principal components. Fig. 4 shows the loadings of PC1 and PC2 of all *m/z* features on both peak matrices and distinguishes between monoisotopes, isotopes and non-annotated ions. The monoisotopic ions tend to have larger loadings on the PCA, indicating that the variance is mainly led by monoisotopic peaks.

Finally, we also analyzed the extent to which monoisotopic peaks influence a segmentation process. To this end, we applied the k-means algorithm to the datasets with all the peaks, and with only the monoisotopic ions. The number of clusters was selected to suit the morphology of the tissues. Fig. 5 shows the results of this procedure. The clusters have the same pixel distribution for both datasets, which indicates that the monoisotopic peaks have a predominant role in establishing the centers of the clustering.

4. Discussion

The annotation of low molecular weight compounds (below 1200 Da) in MSI datasets still has some limitations. As shown in MALDI-TOF annotation results, datasets acquired with TOF mass spectrometers with a resolving power less than 30,000 tend to suffer from overlapping peaks (i.e. isobaric species with very similar mass do not resolve completely). This problem can still arise, although to a much lesser extent, with high resolving power MS analyzers. For instance, if the M+0 peak of a compound A overlaps the M+1 peak of another compound B, the ILS of the M+0 peak of compound B decreases, making it difficult to annotate (supplementary information, section 4). Moreover, when both compounds are co-localized in the same regions of the tissue, the overlapping is harder to detect as peak picking cannot find pixels where both peaks are well resolved. These cases could be addressed with peak deconvolution algorithms, which would split all the isotopic ions from overlapping peaks increasing the scores of peak picking algorithms and generating more annotations. At the same time, peak deconvolution algorithms could benefit from previous peak annotation results by searching for overlapped peaks for which the peak annotation algorithm has previously failed. This would reduce the load of the overall process. As far as we know, no deconvolution algorithms have been reported with this exclusive purpose in the context of MSI, which could be a line of further work. At the end, overlapping peaks is an issue that affects rMSIannotation to some extend and more generally, to all the automatic annotation procedures. We presume that overlapping is one of the reasons why METASPACE encourages users to submit ultra-high-resolution datasets.

Adduct annotation is a problem that is harder to address than isotope annotation. First, there are no general rules applicable to the intensity ratios between M+0 adduct ions, since adduct generation depends on experimental conditions [17]. Some

Table 3
Coinciding annotations of MALDI-FT-ICR dataset 1 between rMSIannotation and METASPACE.

Formula	Adduct	*m/z*	METASPACE (FDR 10%)	METASPACE (FDR 20%)	rMSIannotation
$C_{38}H_{72}O_8S$	$[M+H]^+$	689.502	–	–	x
$C_{40}H_{76}O_8S$	$[M+H]^+$	717.534	–	–	x
$C_{40}H_{77}O_8PS$	$[M+H]^+$	749.515	–	x	–
$C_{46}H_{77}NO_8$	$[M+H]^+$	772.573	–	–	x
$C_{44}H_{76}NO_8P$	$[M+H]^+$	778.538	X	x	x
$C_{47}H_{79}NO_8$	$[M+H]^+$	786.588	–	–	x
$C_{46}H_{76}O_8S$	$[M+H]^+$	789.533	–	x	x
$C_{44}H_{75}O_8PS$	$[M+H]^+$	795.501	X	x	x
$C_{49}H_{92}O_6$	$[M+Na]^+$	799.679	x	x	x
$C_{46}H_{75}NO_{10}$	$[M+H]^+$	802.547	–	–	–
$C_{45}H_{77}O_8PS$	$[M+H]^+$	809.516	x	x	x
$C_{44}H_{87}NO_{10}$	$[M+Na]^+$	812.622	–	–	–
$C_{50}H_{94}O_6$	$[M+Na]^+$	813.695	x	x	–
$C_{48}H_{80}O_8S$	$[M+H]^+$	817.565	–	–	x
$C_{46}H_{79}O_8PS$	$[M+H]^+$	823.533	x	x	x
$C_{45}H_{87}NO_{10}$	$[M+Na]^+$	824.622	–	–	–
$C_{50}H_{83}NO_8$	$[M+H]^+$	826.620	–	–	x
$C_{45}H_{89}NO_{10}$	$[M+Na]^+$	826.640	–	–	–
$C_{51}H_{96}O_6$	$[M+Na]^+$	827.712	x	x	x
$C_{48}H_{82}NO_8P$	$[M+H]^+$	832.585	–	–	x
$C_{50}H_{82}O_8S$	$[M+H]^+$	843.579	–	–	x
$C_{53}H_{90}O_6$	$[M+Na]^+$	845.664	–	x	x
$C_{48}H_{81}O_8PS$	$[M+H]^+$	849.547	x	x	x
$C_{53}H_{98}O_6$	$[M+Na]^+$	853.726	–	x	x
$C_{53}H_{100}O_6$	$[M+Na]^+$	855.746	–	–	–
$C_{52}H_{80}NO_8P$	$[M+H]^+$	878.569	x	x	x
$C_{49}H_{91}NO_{11}$	$[M+Na]^+$	892.649	–	–	–
$C_{50}H_{81}O_8PS$	$[M+H]^+$	895.531	–	x	x
$C_{57}H_{96}O_6$	$[M+Na]^+$	899.712	x	x	x
$C_{54}H_{79}NO_{10}$	$[M+H]^+$	902.578	–	–	x
$C_{59}H_{100}O_6$	$[M+H]^+$	905.759	–	x	x
$C_{57}H_{108}O_6$	$[M+Na]^+$	911.804	–	–	–
$C_{57}H_{96}O_6$	$[M+K]^+$	915.685	–	–	x
$C_{59}H_{100}O_6$	$[M+Na]^+$	927.742	x	x	x
$C_{61}H_{86}O_6$	$[M+Na]^+$	937.634	–	–	x
$C_{61}H_{94}O_6$	$[M+Na]^+$	945.695	–	x	x
$C_{61}H_{96}O_6$	$[M+Na]^+$	947.711	–	x	x
$C_{61}H_{98}O_6$	$[M+Na]^+$	949.727	x	x	x
$C_{61}H_{100}O_6$	$[M+Na]^+$	951.743	–	x	x

compounds tend to ionize better with one specific adduct [32], but this still depends heavily on the sample preparation and the matrix applied. Second, the mass distances between adducts may be similar to mass distances between different compounds or neutral loses. For example, the mass of the ammonium cation is 18.034 Da, which is very close to the mass of a neutral loss of water (18.011 Da). And third, the colocalization of adducts of the same compound can be affected by the natural abundance of the adduct elements, for instance, some structures in the brain tissue have a high intrinsic concentration of potassium which can affect the distribution of potassium adducts across the tissue and their intensity in comparison to other adducts [33]. Therefore, we rely on correlations between adduct ions to assess to likelihood of a set of peaks to originate from the same chemical compound. These limitations result in adduct annotations being less reliable than isotope annotations. To address this, the rMSIannotation strategy consists in presenting to the user all the possible annotations with its scores to facilitate a manually guided confirmation of the results.

In the presented FT-ICR dataset 1, rMSIannotation found more coinciding annotations with the original paper than METASPACE. For the FT-ICR dataset 2, we were able to replicate the previous METASPACE annotations. These results could be attributed to the differences in the isotope annotation criteria. METASPACE annotations are based on isotopic patterns generated using libraries, reducing the possible annotations to the compounds available in those libraries. This limits the annotation of MSI experiment from understudied organisms like microalgae and precludes compound discovery. On the other hand, rMSIannotation measures and validates isotope peaks intensity using intrinsic chemical information, common for all organic compounds, without relying on compound libraries. Moreover, the output of rMSIannotation can be easily integrated in custom R scripts to filter ions and select the non-redundant features to approach the bio-statistical analysis more reliably.

We also investigated the isotope annotation module as a variable selection method by retaining only monoisotopic peaks. The results show that the annotated monoisotopic peaks play a predominant role (i.e. a considerable weight in the loadings) in determining the result of a PCA (Figs. 3 and 4), and in establishing the centers of a common clustering procedure like k-means (Fig. 5). This is probably because monoisotopic peaks have more intensity than their isotopes (this only applies to molecules with fewer than 93 carbon atoms) and that annotated monoisotopic peaks tend to have larger intensities than non-annotated monoisotopic peaks. This suggests that rMSIannotation can annotate the peaks with highest statistical relevance and that in most cases, these peaks are enough to summarize the dataset, which might be something desirable depending on the objectives of the study.

5. Conclusion

We presented rMSIannotation, a software tool that annotates carbon isotopes and adducts for MSI dataset in the low mass range. rMSIannotation is useful for putative identification of compounds

Table 4
Coinciding annotations of MALDI-FT-ICR dataset 2 between rMSIannotation and METASPACE.

Formula	Adduct	*m/z*	rMSIannotation	ILS
$C_{35}H_{66}O_4$	$[M+H]^+$	551.503	x	0.977
$C_{37}H68O_4$	$[M+H]^+$	577.519	x	0.975
$C_{28}H_{33}O_{14}$	$[M+Na]^+$	616.176	x	0.988
$C_{37}H_{71}O_8P$	$[M+Na]^+$	697.478	x	0.938
$C_{37}H_{71}O_8P$	$[M+K]^+$	713.452	x	0.965
$C_{39}H_{73}O_8P$	$[M+Na]^+$	723.494	x	0.957
$C_{41}H_{83}N_2O_6P$	$[M+H]^+$	731.606	x	0.909
$C_{40}H_{80}NO_8P$	$[M+H]^+$	734.569	x	0.988
$C_{39}H_{73}O_8P$	$[M+K]^+$	739.468	x	0.951
$C_{39}H_{79}N_2O_6P$	$[M+K]^+$	741.531	x	0.963
$C_{41}H_{82}NO_8P$	$[M+H]^+$	748.585	x	0.979
$C_{41}H_{83}N_2O_6P$	$[M+Na]^+$	753.588	x	0.976
$C_{40}H_{80}NO_8P$	$[M+Na]^+$	756.551	x	0.977
$C_{42}H_{84}NO_8P$	$[M+H]^+$	762.601	x	0.962
$C_{41}H_{83}N_2O_6P$	$[M+K]^+$	769.562	x	0.982
$C_{43}H_{74}NO_7P$	$[M+Na]^+$	770.510	x	0.912
$C_{40}H_{78}NO_8P$	$[M+K]^+$	770.510	x (isobaric)	0.912
$C_{43}H_{76}NO_7P$	$[M+Na]^+$	772.525	x	0.967
$C_{40}H_{80}NO_8P$	$[M+K]^+$	772.525	x (isobaric)	0.967
$C_{42}H_{84}NO_8P$	$[M+Na]^+$	784.583	x	0.924
$C_{45}H_{76}NO_7P$	$[M+Na]^+$	796.525	x	0.896
$C_{42}H_{80}NO_8P$	$[M+K]^+$	796.525	x (isobaric)	0.896
$C_{45}H_{80}NO_7P$	$[M+Na]^+$	800.557	x	0.797
$C_{42}H_{84}NO_8P$	$[M+K]^+$	800.557	x (isobaric)	0.797
$C_{43}H_{78}NO_8P$	$[M+K]^+$	806.510	x	0.921
$C_{44}H_{80}NO_8P$	$[M+K]^+$	820.525	x	0.968
$C_{44}H_{84}NO_8P$	$[M+K]^+$	824.557	x	0.950
$C_{44}H_{86}NO_8P$	$[M+K]^+$	826.572	x	0.970
$C_{45}H_{78}NO_8P$	$[M+K]^+$	830.510	x	0.942
$C_{46}H_{84}NO_8P$	$[M+Na]^+$	832.583	x	0.937
$C_{46}H_{80}NO_8P$	$[M+K]^+$	844.525	x	0.968
$C_{46}H_{82}NO_8P$	$[M+K]^+$	846.541	x	0.919
$C_{46}H_{84}NO_8P$	$[M+K]^+$	848.557	x	0.958
$C_{48}H_{91}NO_8$	$[M+K]^+$	848.638	x	0.959
$C_{48}H_{84}NO_8P$	$[M+K]^+$	872.557	x	0.933

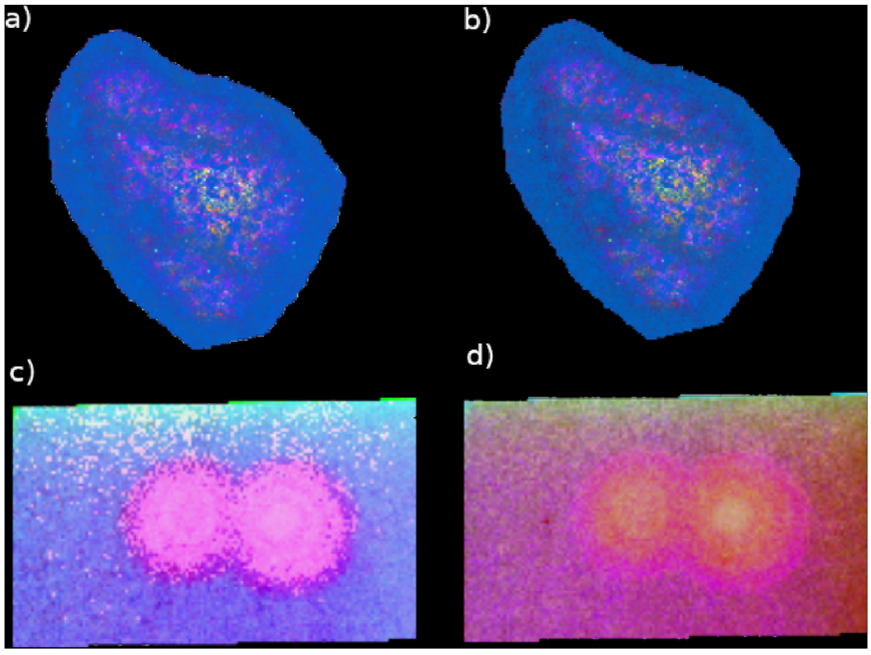


Fig. 3. Representation of the first three principal components on the tissue in RGB. Red channel for PC1, green channel for PC2 and blue channel for PC3. a) TOF dataset with all the peaks. b) TOF dataset with only annotated monoisotopic peaks. c) FT-ICR dataset with all the peaks. d) FT-ICR with only annotated monoisotopic peaks. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

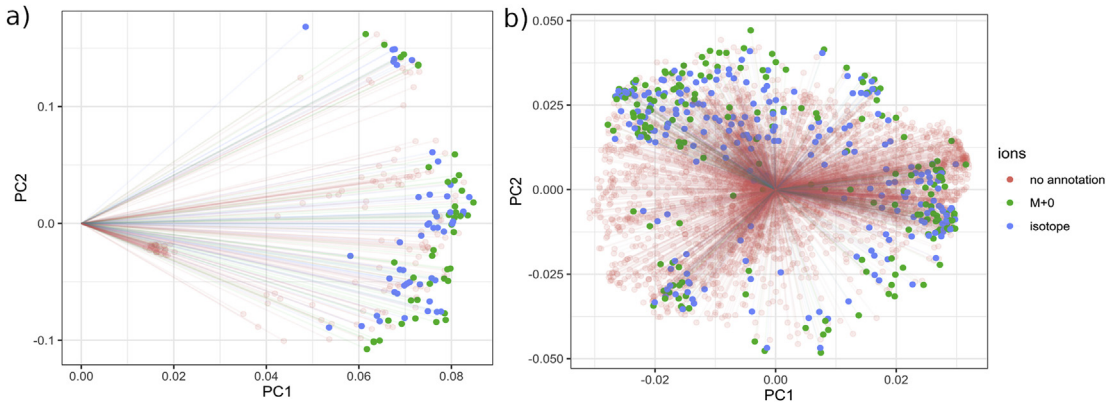


Fig. 4. a) Loadings of PC1 and PC2 of the TOF dataset b) Loadings of PC1 and PC2 of the FT-ICR datasets. Every point in the graphs represents an *m*/*z* feature in the datasets. Green points represent the peaks annotated as monoisotopic, blue points are peaks annotated as isotopes (M+1, M+2, etc.) and red points are peaks that have not been annotated by rMSIannotation. (For interpretation of the references to color in this figure legend, the reader is referred to the Web version of this article.)

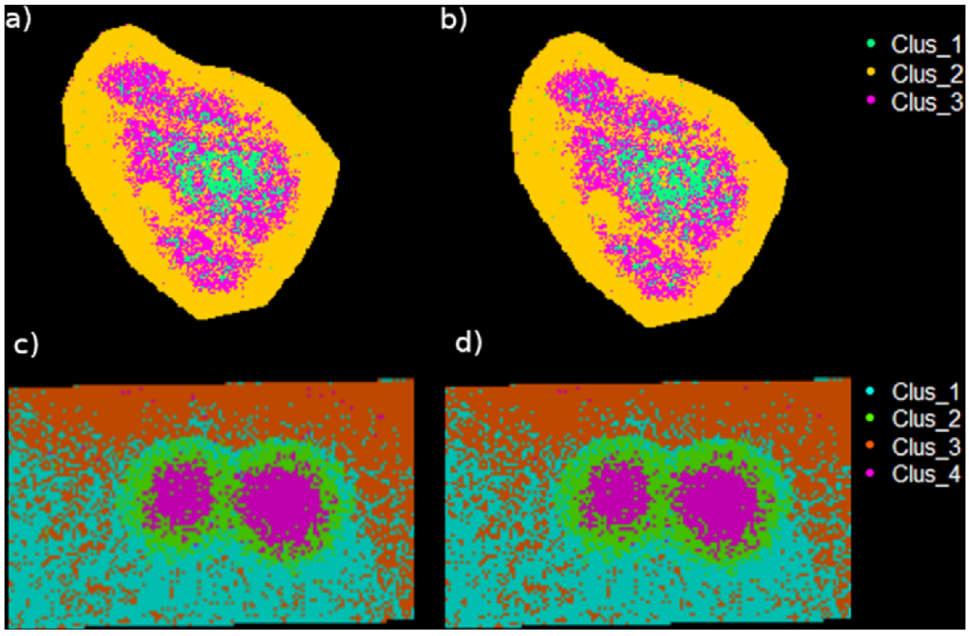


Fig. 5. a) k-means clustering of the TOF dataset with all peaks with k = 3. b) k-means clustering of the TOF dataset with only the monoisotopic ions with k = 3. c) k-means clustering of the FT-ICR dataset with all peaks with k = 4. d) k-means clustering of the FT-ICR dataset with only the monoisotopic ions with k = 4.

and variable reduction strategies; and can be integrated in any low-weight compounds MSI data analysis workflows. The results show that rMSIannotation automatically extracts valuable information from both high (TOF) and ultra-high (FT-ICR) resolution spectrometers. The presented algorithm demonstrated a high performance and annotation confidence when compared to the established metabolomics MSI annotation platform: METASPACE and to the manual annotation approaches.

The tool is integrated into the MSI processing R package rMSIproc <https://github.com/prafols/rMSIproc>, which processes and annotates data within the same software environment. This expands the possibilities of MSI data analysis for biological research by reducing data processing and manual inspection time.

CRediT authorship contribution statement

Lluc Sementé: Conceptualization, Methodology, Software, Validation, Writing – original draft, Visualization. **Gerard Baquer:** Writing – review & editing. **María García-Altares:** Writing – review & editing, Conceptualization. **Xavier Correig-Blanchar:** Writing – review & editing, Funding acquisition. **Pere Ràfols:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

The authors acknowledge the financial support of the Spanish Ministry of Economy and Competitivity through project RTI2018-096061-B-100. LS acknowledges the financial support of Universitat Rovira i Virgili through the pre-doctoral grant 2017PMF-PIPF-60. GB acknowledges the financial support of the European

Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie grant agreement No. 713679 and the Universitat Rovira i Virgili (URV). MGA acknowledges the financial support from the Agency for Management of University and Research Grants of the Generalitat de Catalunya (AGAUR) through the post-doctoral grant 2018 BP 00188.

The authors thank Svetlana Uzbekova from French National Institute for Agriculture, Food, and Environment (INRAE) for kindly sharing the MALDI-TOF dataset upon request and to Guy Schleyer from The Vardi group for his useful comments on the manuscript and sharing the FT-ICR dataset.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.aca.2021.338669.

References

[1] L.A. McDonnell, R.M.A. Heeren, Imaging mass spectrometry, Mass Spectrom. Rev. 26 (2007) 606–643, https://doi.org/10.1002/mas.

[2] T.C. Rohner, D. Staab, M. Stoeckli, MALDI mass spectrometric imaging of biological tissue sections, Mech. Ageing Dev. 126 (1) (2005) 177–185.

[3] K. Chughtai, R.M.A. Heeren, Mass spectrometric imaging for biomedical tissue analysis, Chem. Rev. 110 (5) (2010) 3237–3277, https://doi.org/10.1021/cr100012c.

[4] J.L. Norris, D.S. Cornett, J.A. Mobley, M. Andersson, E.H. Seeley, P. Chaurand, R.M. Caprioli, Processing MALDI mass spectra to improve mass spectral direct tissue analysis, Int. J. Mass Spectrom. 260 (2) (2007) 212–221.

[5] L.A. McDonnell, A. van Remoortere, R.J.M. van Zeijl, A.M. Deelder, Mass spectrometry image correlation: quantifying colocalization, J. Proteome Res. 7 (8) (2008) 3619–3627, https://doi.org/10.1021/pr800214d.

[6] P. Ràfols, D. Vilalta, J. Brezmes, N. Cañellas, E. del Castillo, O. Yanes, N. Ramírez, X. Correig, Signal preprocessing, multivariate analysis and software tools for MA(LDI)-TOF mass spectrometry imaging for biological applications, Mass Spectrom. Rev. 37 (3) (2018) 281–306, https://doi.org/10.1002/mas.21527.

[7] T. Alexandrov, MALDI imaging mass spectrometry: statistical data analysis and current computational challenges, BMC Bioinf. 13 (16) (2012) S11, https://doi.org/10.1186/1471-2105-13-S16-S11.

[8] E. del Castillo, L. Sementé, S. Torres, P. Ràfols, N. Ramírez, M. Martins-Green, M. Santafe, X. Correig, RMsikeyion: an ion filtering r package for untargeted analysis of metabolomic LDI-MS images, Metabolites 9 (8) (2019), https://doi.org/10.3390/metabo9080162.

[9] S.A. Thomas, A.M. Race, R.T. Steven, I.S. Gilmore, J. Bunch, Dimensionality reduction of mass spectrometry imaging data using autoencoders, IEEE Symposium Series on Computational Intelligence (SSCI), 2016, pp. 1–7, 2016.

[10] L.A. McDonnell, A. van Remoortere, N. de Velde, R.J.M. van Zeijl, A.M. Deeldera, Imaging mass spectrometry data reduction: automated feature identification and extraction, J. Am. Soc. Mass Spectrom. 21 (12) (2010) 1969–1978, https://doi.org/10.1021/jasms.8b03661.

[11] C. Kuhl, R. Tautenhahn, C. Böttcher, T.R. Larson, S. Neumann, CAMERA: an integrated strategy for compound spectra extraction and annotation of liquid chromatography/mass spectrometry data sets, Anal. Chem. 84 (1) (2012) 283–289, https://doi.org/10.1021/ac202450g.

[12] L. Wang, X. Xing, L. Chen, L. Yang, X. Su, H. Rabitz, W. Lu, J.D. Rabinowitz, Peak annotation and verification engine for untargeted LC–MS metabolomics, Anal. Chem. 91 (3) (2019) 1838–1846, https://doi.org/10.1021/acs.analchem.8b03132.

[13] X. Domingo-Almenara, J.R. Montenegro-Burke, H.P. Benton, G. Annotation Siuzdak, A computational solution for streamlining metabolomics analysis, Anal. Chem. 90 (1) (2018) 480–489, https://doi.org/10.1021/acs.analchem.7b03929.

[14] C.A. Hughey, C.L. Hendrickson, R.P. Rodgers, A.G. Marshall, K. Qian, Kendrick mass defect spectrum: a compact visual analysis for ultrahigh-resolution broadband mass spectra, Anal. Chem. 73 (19) (2001) 4676–4681, https://doi.org/10.1021/ac010560w.

[15] L.A. Lerno, J.B. German, C.B. Lebrilla, Method for the identification of lipid classes based on referenced Kendrick mass analysis, Anal. Chem. 82 (10) (2010) 4236–4245, https://doi.org/10.1021/ac100556g.

[16] L.W. Sumner, A. Amberg, D. Barrett, M.H. Beale, R. Beger, C.A. Daykin, T.W.-M. Fan, O. Fiehn, R. Goodacre, J.L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A.N. Lane, J.C. Lindon, P. Marriott, A.W. Nicholls, M.D. Reily, J.J. Thaden, M.R. Viant, Proposed minimum reporting standards for chemical analysis chemical analysis working group (CAWG) metabolomics standards initiative (MSI), Metabolomics: Off. J. Metabol. Soc. 3 (3) (2007) 211–221, https://doi.org/10.1007/s11306-007-0082-2.

[17] O. Senan, A. Aguilar-Mogas, M. Navarro, J. Capellades, L. Noon, D. Burks, O. Yanes, R. Guimerà, M.CliqueMS. Sales-Pardo, A computational tool for annotating in-source metabolite ions from LC-MS untargeted metabolomics data based on a coelution similarity network, Bioinformatics 35 (20) (2019) 4089–4097, https://doi.org/10.1093/bioinformatics/btz207.

[18] A. Alonso, A. Julià, A. Beltran, M. Vinaixa, M. Díaz, L. Ibañez, X. Correig, S. Marsal, AStream: an R package for annotating LC/MS metabolomic data, Bioinformatics 27 (9) (2011) 1339–1340, https://doi.org/10.1093/bioinformatics/btr138.

[19] N.J. Bond, A. Koulman, J.L. Griffin, Z. MassPix Hall, An R Package for Annotation and Interpretation of Mass Spectrometry Imaging Data for Lipidomics, Metabolomics (2017), https://doi.org/10.1007/s11306-017-1252-5.

[20] A. Palmer, P. Phapale, I. Chernyavsky, R. Lavigne, D. Fay, A. Tarasov, V. Kovalev, J. Fuchser, S. Nikolenko, C. Pineau, M. Becker, T. Alexandrov, FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry, Nat. Methods 14 (1) (2017) 57–60, https://doi.org/10.1038/nmeth.4072.

[21] L. Reiter, M. Claassen, S.P. Schrimpf, M. Jovanovic, A. Schmidt, J.M. Buhmann, M.O. Hengartner, R. Aebersold, Protein identification false discovery rates for very large proteomics data sets generated by tandem mass spectrometry, Mol. Cell. Proteom. 8 (11) (2009) 2405, https://doi.org/10.1074/mcp.M900317-MCP200. LP – 2417.

[22] P. Ràfols, B. Heijs, E. del Castillo, O. Yanes, L.A. McDonnell, J. Brezmes, I. Pérez-Taboada, M. Vallejo, M. García-Altares, X. Correig, RMSIproc: an R package for mass spectrometry imaging data processing, Bioinformatics 36 (11) (2020) 3618–3619, https://doi.org/10.1093/bioinformatics/btaa142.

[23] P.S. Bertevello, A.P. Teixeira-Gomes, A. Seyer, A.V. Carvalho, V. Labas, M.C. Blache, C. Banliat, L.A.V. Cordeiro, V. Duranthon, P. Papillier, V. Maillard, S. Elis, S. Uzbekova, Lipid identification and transcriptional analysis of controlling enzymes in bovine ovarian follicle, Int. J. Mol. Sci. 19 (10) (2018), https://doi.org/10.3390/ijms19103261.

[24] G. Schleyer, N. Shahaf, C. Ziv, Y. Dong, R.A. Meoded, E.J.N. Helfrich, D. Schatz, S. Rosenwasser, I. Rogachev, A. Aharoni, J. Piel, A. Vardi, In plaque-mass spectrometry imaging of a bloom-forming alga during viral infection reveals a metabolic shift towards odd-chain fatty acid lipids, Nat. Microbiol. 4 (3) (2019) 527–538, https://doi.org/10.1038/s41564-018-0336-y.

[25] K. Haug, K. Cochrane, V.C. Nainala, M. Williams, J. Chang, K.V. Jayaseelan, C. O'Donovan, MetaboLights: a resource evolving in response to the needs of its scientific community, Nucleic Acids Res. 48 (D1) (2019) D440–D444, https://doi.org/10.1093/nar/gkz1019.

[26] D.S. Wishart, Y.D. Feunang, A. Marcu, A.C. Guo, K. Liang, R. Vázquez-Fresno, T. Sajed, D. Johnson, C. Li, N. Karu, Z. Sayeeda, E. Lo, N. Assempour, M. Berjanskii, S. Singhal, D. Arndt, Y. Liang, H. Badran, J. Grant, A. Serra-Cayuela, Y. Liu, R. Mandal, V. Neveu, A. Pon, C. Knox, M. Wilson, C. Manach, A. Scalbert, HMDB 4.0: the human metabolome database for 2018, Nucleic Acids Res. 46 (D1) (2017) D608–D617, https://doi.org/10.1093/nar/gkx1089.

[27] E. Fahy, M. Sud, D. Cotter, S. Subramaniam, LIPID MAPS online tools for lipid research, Nucleic Acids Res. 35 (suppl_2) (2007) W606–W612, https://doi.org/10.1093/nar/gkm324.

[28] C. Guijas, J.R. Montenegro-Burke, X. Domingo-Almenara, A. Palermo, B. Warth, G. Hermann, G. Koellensperger, T. Huan, W. Uritboonthai, A.E. Aisporna, D.W. Wolan, M.E. Spilker, H.P. Benton, G. Siuzdak, METLIN: a technology platform for identifying knowns and unknowns, Anal. Chem. 90 (5) (2018) 3156–3164, https://doi.org/10.1021/acs.analchem.7b04424.

[29] C.R.C. Press, Taylor, Francis Group, an I. G. Company 2020 (P2). Dictionary of Natural Products 29.1, Chemical Search, http://dnp.chemnetbase.com/faces/chemical/ChemicalSearch.xhtml. (Accessed 22 July 2020).

[30] J.M. Fulton, B.J. Kendrick, G.R. DiTullio, B.A.S. van Mooy, Alkenone unsaturation during virus infection of emiliania huxleyi, Org. Geochem. 111 (2017) 82–85.

[31] C.A. Llewellyn, C. Evans, R.L. Airs, I. Cook, N. Bale, W.H. Wilson, The response of carotenoids and chlorophylls during virus infection of emiliania huxleyi (prymnesiophyceae), J. Exp. Mar. Biol. Ecol. 344 (1) (2007) 101–112.

[32] J. Garate, S. Lage, L. Martín-Saiz, A. Perez-Valle, B. Ochoa, M.D. Boyano, R. Fernández, J.A. Fernández, Influence of lipid fragmentation in the data analysis of imaging mass spectrometry experiments, J. Am. Soc. Mass Spectrom. 31 (3) (2020) 517–526, https://doi.org/10.1021/jasms.9b00090.

[33] J.A. Hankin, S.E. Farias, R.M. Barkley, K. Heidenreich, L.C. Frey, K. Hamazaki, H.-Y. Kim, R.C. Murphy, MALDI mass spectrometric imaging of lipids in rat brain injury models, J. Am. Soc. Mass Spectrom. 22 (6) (2011), https://doi.org/10.1021/jasms.8b04045.