A robust data processing and normalization strategy for MALDI mass spectrometric imaging

Judith M. Fonville[a], Claire Carter[b], Olivier Cloarec[a,c], Jeremy K. Nicholson[a], John C. Lindon[a], Josephine Bunch[b] and Elaine Holmes[a*]*

a) Biomolecular Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, Sir Alexander Fleming Building, South Kensington, London SW7 2AZ, United Kingdom.

b) School of chemistry, University of Birmingham, Edgbaston, Birmingham B15 2TT, United Kingdom.

c) Korrigan Sciences Ltd., Imperial Place, Maidenhead SL6 2GN, United Kingdom.

* To whom correspondence should be addressed.

Josephine Bunch:

Tel: +44(0)121418810; Fax: +44(0)121414403; Email: j.bunch@bham.ac.uk

Elaine Holmes:

Tel: +44(0)2075943220; Fax: +44(0)2075943226; Email: elaine.holmes@imperial.ac.uk

ABSTRACT: Matrix-assisted laser desorption/ionization (MALDI) mass spectrometry imaging (MSI) provides localized information about the molecular content of a tissue sample. To derive reliable conclusions about MSI data, it is necessary to implement appropriate processing steps in order to compare peak intensities across the different pixels comprising the image. Here, we review commonly used normalization methods, and propose a rational data processing strategy, for robust evaluation and modeling of MSI data. The approach includes newly developed heuristic methods for selecting biologically relevant peaks and pixels to reduce the size of a data set and remove the influence of the applied MALDI matrix. The methods are demonstrated on a MALDI MSI data set of a sagittal section of rat brain (4750 bins, m/z = 50-1000, 111 × 185 pixels) and the proposed preferred normalization method uses the median intensity of selected peaks, which were determined to be independent of the MALDI matrix. This was found to effectively compensate for a range of known limitations associated with the MALDI process and irregularities in MS image sampling routines. This new approach is relevant for processing of all MALDI MSI data sets, and thus likely to have impact in biomarker profiling, preclinical drug distribution studies, and studies addressing underlying molecular mechanisms of tissue pathology.

KEYWORDS: mass spectrometry imaging, processing, normalization, peak selection, total ion current, TIC, matrix, MALDI, MSI, principal component analysis, lipids, rat brain.

INTRODUCTION

Molecular profiling methods such as mass spectrometry imaging (MSI) can be used to study topographic distributions of molecules. MSI generates a series of mass spectra, which are obtained from sequential locations forming a grid, covering the topology of a tissue slice.[1-5] Thus, these high-dimensional data can be interpreted as a full mass spectrum at a given spatial point (pixel), providing a molecular fingerprint for different spatially resolved regions, or as a map of a given ion abundance over a two-dimensional set of pixels, providing localized information for different biomolecules. Such localized molecular information provides the opportunity to directly investigate the link between tissue structure and function.

Direct analysis of tissue with the soft ionization technique of matrix-assisted laser desorption/ionization (MALDI) mass spectrometry (MS) has been used to study the distribution of proteins,[1, 6-8] lipids,[9-10] metabolites[11-12] and drugs,[13-14] and has been performed on a large number of sample types,[1, 15-16] such as plants,[12] bacterial colonies,[17] drug-treated tissue,[13] and human surgical specimens such as cancerous tissue in brain, breast and ovaries.[6, 8, 18] Whole animals sections have been imaged as part of preclinical drug evaluations.[14] Data are typically processed and evaluated with standard MSI analysis software, such as BIOMAP[19], ClinProTools[20] or other such packages.[21] The multivariate nature of MALDI MSI data sets has been addressed using approaches such as principal component analysis (PCA), which has been performed on selected MSI pixels[6, 22] or the complete sample.[23]

To retain maximum biological information, it is preferable to use all available pixels and *m/z* peaks. However, a side effect of the MALDI process is that the applied matrix is detected as peaks that interfere with the spectral profile, especially for *m/z* regions <1000. Evaluating the images of *a priori* determined *m/z* values is a commonly used approach, but this precludes identification of novel biomarkers for a given tissue type or lesion. Here, we propose a robust biological interpretation of these data by including all data, except pixels that are outside the sample border and *m/z* peaks that are

associated with the applied matrix.

In addition to biological variation, several instrumental and experimental influences introduce systematic and random variations in signal intensity of MSI data sets. One commonly observed effect is a varying level of spectral response intensity, which affects image intensities. This is mainly caused by the inherent limitations of the MALDI technique and sample preparation protocols, such as uneven MALDI matrix coating, differential ionization efficiencies and crystal inhomogeneity.[16, 24-25] Normalization is commonly applied to address this issue, in order to transform measured intensities to a comparable scale, without altering the biological information. The normalization techniques applied to MSI data sets are often limited to basic procedures implemented in the commonly used software, which force all pixels to have equal total ion current (TIC)[24], possibly excluding a high intensity signal, or normalization through scaling peak intensities by a factor calculated from a reference molecule, such as a matrix signal.[12] These normalization techniques rely on a set of assumptions that may not be fulfilled, for example the fact that molecular composition changes in different tissues is not generally taken into account. Recently, normalization procedures that use the level of noise to normalize the signal intensity were reported for protein species, and the benefits and necessity of data normalization were impressively illustrated.[26] Unfortunately, the suggested methods may be difficult to implement if apparent spectral noise is relatively low or absent, for example, in the analysis of phospholipids using a Qstar instrument platform (which may be matrix and solvent system dependent), and in cases where data are binned or where many signals are found in the *m/z*-range, reducing the opportunity for selecting *m/z*-variables representing noise only, and no molecular signal.

Although it is clear that processing these highly complex imaging data will greatly influence the results of subsequent analyses,[24, 27] there is still little understanding of the optimal approach. For example, the presence of heteroscedastic noise can adversely affect standard statistical and pattern recognition tools, however, variance stabilizing transformations are not commonly employed in the current MSI literature. Here, we propose a rationalized data processing strategy that includes heuristic

methods for selecting biologically relevant peaks and pixels to reduce the size and complexity of MSI data sets. Different normalization techniques were compared, and the underlying assumptions are discussed. An intuitive normalization based on the median pixel intensity is suggested, and the final processing scheme is demonstrated with a formalin fixed sagittal rat brain section, acquired with spatial resolution 100 μm × 100 μm and with *m/z* range of 50-1000.

MATERIALS AND METHODS

MALDI MS imaging of formalin fixed rat brain was recently described.[28] Briefly, a whole rat brain was fixed for 72 hrs, frozen to -20°C and bisected along the midline. Sagittal sections were taken at 12 μm using a Leica cryostat and thaw mounted onto stainless steel MALDI target plates. Plates were coated in 5 mg/mL *Alpha*-cyano-4-hydroxy cinnamic acid (αCHCA) matrix material, prepared in 80% methanol (0.1% trifluoracetic acid) using an automated matrix deposition system (TM sprayer from HTX Technologies, NC, USA). Plates were sprayed (8 cycle repeats) at 150 ºC, 10 psi, a flow rate of 0.25 mL/min with a stage velocity of 500 mm/min. Mass spectrometry imaging was carried out on a QqTOF (Qstar Elite) mass spectrometer (Applied Biosystems, Foster City, USA), operated in positive ion reflectron mode. An Nd:YAG (355 nm) laser was operated at 20% available power (2.1 μJ) with a repetition rate of 500 Hz. Sequential spectra were acquired at a resolution of 100 μm^2 using the `dynamic pixel' setting (oMALDI, 5.1).

The data from this experiment were imported into MATLAB (Natick, USA) with a bin size of $\Delta m/z$ = 0.2, and "unfolded" to form an array where each row represents a pixel and the columns represent the different *m/z* bins. The effects of the different normalization approaches were evaluated by showing single *m/z* images, and by evaluating the principal component analysis (PCA) scores for each pixel. Full details of the materials used, sample preparation and mass spectrometry technique are given in the *Supporting Information*. The data preprocessing and multivariate statistical methodology are also given there.

RESULTS AND DISCUSSION

Mass spectra were acquired for the range *m/z* 50-1000 at 100 μm^2, thus the initial data set consisted of integrated signal intensities for 20535 pixels in 4750 bins. We describe an evaluation of predate processing steps for efficient handling and interpretation of this large and complex data set.

Peak selection

Peak selection decreases the number of non-informative peaks, reduces the data set size and subsequent modeling times, and has been shown to be necessary to obtaining useful multivariate models for MSI data.[18, 29] However, the quality and quantity of spectral information varies between pixels, and thus a simple criterion for rejection of noisy variables such as signal-to-noise ratio is not directly applicable. For this data set we consecutively apply two pragmatic methods, based on the localization of matrix peak signals to remove matrix-related peaks, and the presence of anatomical structure to remove non-anatomically distributed peaks, which resulted in a systematic peak selection, retaining ~10% of the original data.

Approach A for peak selection: the use of matrix peaks to remove non-biological variables

Preliminary data analysis demonstrated that some peaks were more prominent in the region outside the tissue sample than within the sample boundaries, a significant number of which arise from the matrix compound. These peaks, although possibly informative, do not directly convey information on biological localization of endogenous molecules; hence these peaks were removed, based on the correlation to the matrix-related peaks (described in the *Supporting Information*). Ions that had a negative correlation with matrix-related peak intensities were retained. It is clear from examples in the different correlation regions, shown in *Figure 1 A*, that positive correlations indeed correspond to variables with a higher signal intensity outside the tissue region than within the tissue, and are therefore unlikely to be biologically relevant (e.g. *m/z* 212.0, $[\alpha CHCA + Na]^+$), whereas peaks with a negative

overall correlation (e.g. *m/z* 769.4), display an anatomically relevant distribution. Interestingly, this analysis increases the current understanding of the contribution from the MALDI matrix solution to the acquired MSI spectra. Typically, a short list of matrix peaks would be identified by researchers and removed prior to analysis, yet in this study 3526 out of 4750 bins were found to be correlated with the matrix distribution patterns. This shows that many more *m/z* values in the data set may be associated with the matrix than are usually considered. It should be noted that some peaks removed with this approach could be from molecules that were delocalized as a result of the matrix application process (in particular if no signal was retained within the sample boundary). Exclusion of these peaks seems reasonable, because they would not convey information on the localization of the molecule in tissue anymore, however, identification of this phenomenon and the corresponding ions would be important if one was specifically interested in these molecules or would like to optimize the sample preparation process. Manual inclusion of peaks is, of course, possible, but the severe delocalization may give results that are difficult to interpret and potentially less biologically meaningful.

Approach B for peak selection: using image anatomy to find relevant variables

The peak selection resulting from approach **A** could be improved for this data set, since noisy variables were still included in the resulting selection. Although many multivariate modeling approaches are able to cope with noisy variables, the model strength decreases with a large number of non-informative variables. For univariate methods, the effects of noise can be even more problematic. Therefore, identification of the variables retained in the selection of approach **A** that lack any relation to anatomy was proposed as a second filter. An indicator of image-related intensity distribution of the variable was defined as the percentage of variance explained (VE) in the first singular value of a singular value decomposition made for each *m/z* image. If high intensity regions for an *m/z* image are randomly distributed, the variance explained by the first singular value will be low, e.g. for *m/z* 407.2, see *Figure 1 B*. On the other hand, if there is any structure in the image, more variance is modeled, as is

shown for *m/z* 873.6 in *Figure 1 B* (more details can be found in the *Supporting Information*).

A "*VE*-threshold" was developed as a pragmatic and user-friendly cut-off for the variable selection approach **B**, which is a heuristic parameter based on pareto-efficiency considerations, calculated as the sum of all explained variances divided by the number of variables (which was 1224, as we only evaluate the variables that were selected with approach **A**). Variables for which the explained variance in the first singular value was higher than this *VE*-threshold of 24.3%, were retained and colored green in *Figure 1 B*. It should be noted that the *VE*-threshold percentage will vary for different data sets, as the levels of variance explained and number of variables change. A high level of explained variance would also be obtained for *m/z* images that are high in the surrounding and low in the sample region, such as the matrix peaks. Therefore, it is necessary to conduct variable selection with approach **A** prior to peak selection with approach **B**. The research question will determine whether only peak selection approach A or both A and B is more appropriate, depending on the relevance of biologically real peaks without localization; additionally, it is good practice for the user to inspect the images of peaks that were removed as a quality control step.

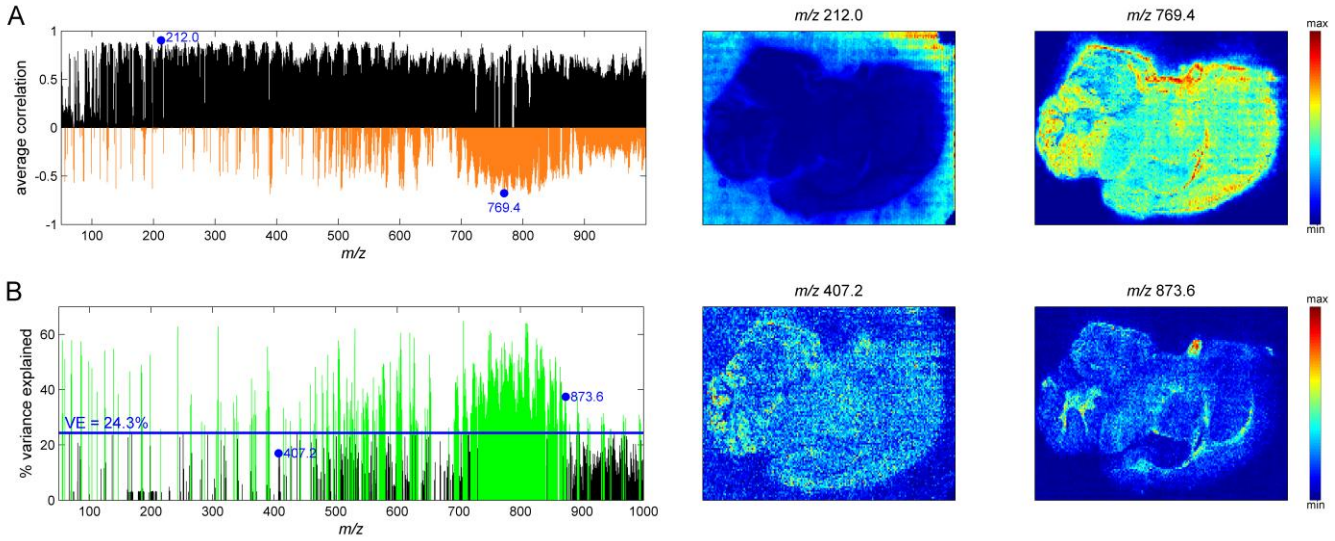


Figure 1. Peak selection process of the sagittal rat brain MALDI MSI data. (A) The average correlation of each variable with the 10 selected matrix related peaks (*Figure S-1 A*) is shown. Retained variables are shown in orange. Images of 2 selected variables demonstrate that positive correlations correspond to higher intensity outside the sample (*m/z* 212.0) and negative correlations show more relevance to the biological tissue (*m/z* 769.4). (B) The variance explained (VE) by the first singular value of a singular value decomposition of the image for each variable is plotted. Variables with lower levels of explained variance (e.g. *m/z* 407.2) conveyed less anatomical relevance. The *VE*-threshold was calculated as 24.3%, and values above this *VE*-threshold are retained (shown in green, e.g. *m/z* 873.6).

Pixel selection

For image interpretation and statistical analyses, it is beneficial to discard pixels that clearly do not arise from sample regions, and we based this selection on the peak intensity profiles of pixels. A fast method to select the sample pixels was found based on the ratio of the total ion current (TIC) of the informative peaks (selected after variable selection with approaches **A** and **B**, 564 variables) versus the TIC of all original 4750 variables. The TIC is the sum of the peak intensity values for a pixel, and the ratio of the TICs based on the full and selected peak lists is shown for all pixels in *Figure 2 A*. It is observed that pixels with a higher TIC from selected variables (TIC_{564}) compared to the total TIC (TIC_{4750}), colored red, correspond to sample pixels. Application of a manually chosen threshold $\log_{10}(TIC_{564}/TIC_{4750}) > -0.5$ results in the selection of the pixels colored green in *Figure 2 B*. To create a fully continuous set of pixels, pixels were included in the final data set if 6 or more pixels in the enclosed 3 × 3 pixel region had been selected with the TIC-ratio threshold. The final pixel selection (13176 out of 20535) is shown in *Figure 2 C*.

The sizes of the data set after the different processing steps are listed in *Table S-1*. The reduction of data set size decreases computational cost and increases model quality and therefore interpretability.[30] We advise users to investigate the signal outside the selected sample region to detect delocalization of signal and to evaluate the matrix application step, prior to extensive data processing and modeling efforts, to ensure data quality. Some instruments allow for a non-rectangular, closer outline of the tissue section to be imaged, and although this may be time-saving, it potentially endangers the opportunity to investigate distant pixels to determine the matrix-related peaks and perform unbiased quality control.

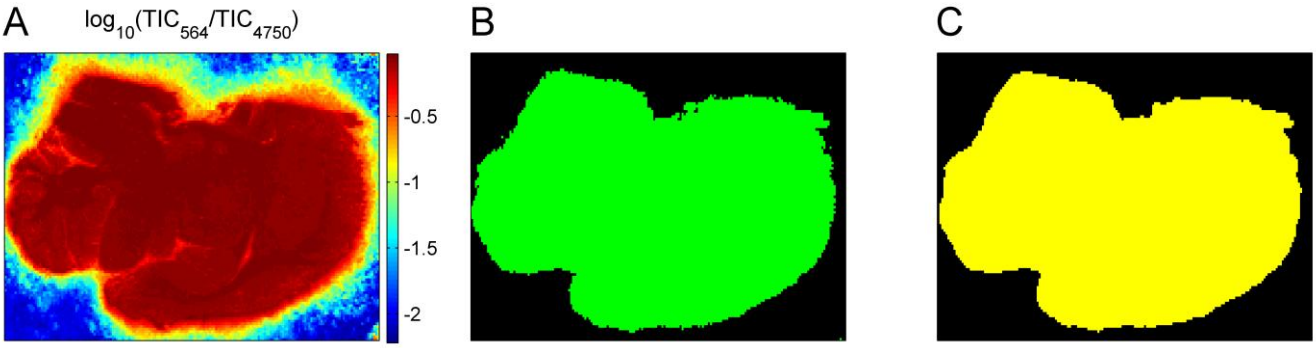


Figure 2. Pixel selection based on TIC. (A) A relatively high value of the total ion current (TIC) from the 564 selected variables (TIC_{564}) versus the TIC from all 4750 variables (TIC_{4750}) corresponds to pixels from the sample, as shown in this image colored with $\log_{10}(TIC_{564}/TIC_{4750})$. (B) A threshold of $\log_{10}(TIC_{564}/TIC_{4750}) > -0.5$ results in the selection of the green-colored pixels. (C) The final continuous pixel selection.

Normalization

Normalization is a procedure that adjusts the individual mass spectra with the goal of making pixels in an MSI image comparable: a correction is made for an experimental bias that would result in higher (or lower) signal intensities in some pixels than in others. Normalization is necessary due to factors such as matrix heterogeneity and application, differential ionization efficiency, ion suppression due to varying sample complexity and composition, instrumental variation and varying levels of analyte solubility and extraction from different tissue regions.[10, 16, 24-26] Normalization correction should address analytical and technical variation, rather than biological variation of interest, and is widely accepted to be necessary for analysis of MSI data, especially to correct for differences in matrix solution coverage and efficiency of analyte incorporation in MALDI.[31] Commonly used methods of normalization in mass spectrometry and their influence on the identification of biomarkers in a non-image setting have been described elsewhere.[32-35]

In the present study, seven methods of normalizing the spectral intensities of each pixel were evaluated. The factor by which the peak intensities of each individual pixel would be divided for each of these normalization methods is shown in *Figure 3*, where 'informative peaks' refers to peaks that remained after variable selection with approaches **A** and **B**. Histogram matching is the only method that does not involve the division by a scalar normalization factor, and could therefore not be displayed. It is clear from *Figure 3* that the different normalization approaches deliver drastically different results, and subsequent data analysis is greatly dependent on an appropriate choice of normalization method. The assumptions underlying each of the normalization methods are fully discussed in the *Supporting Information*, and are summarized here.

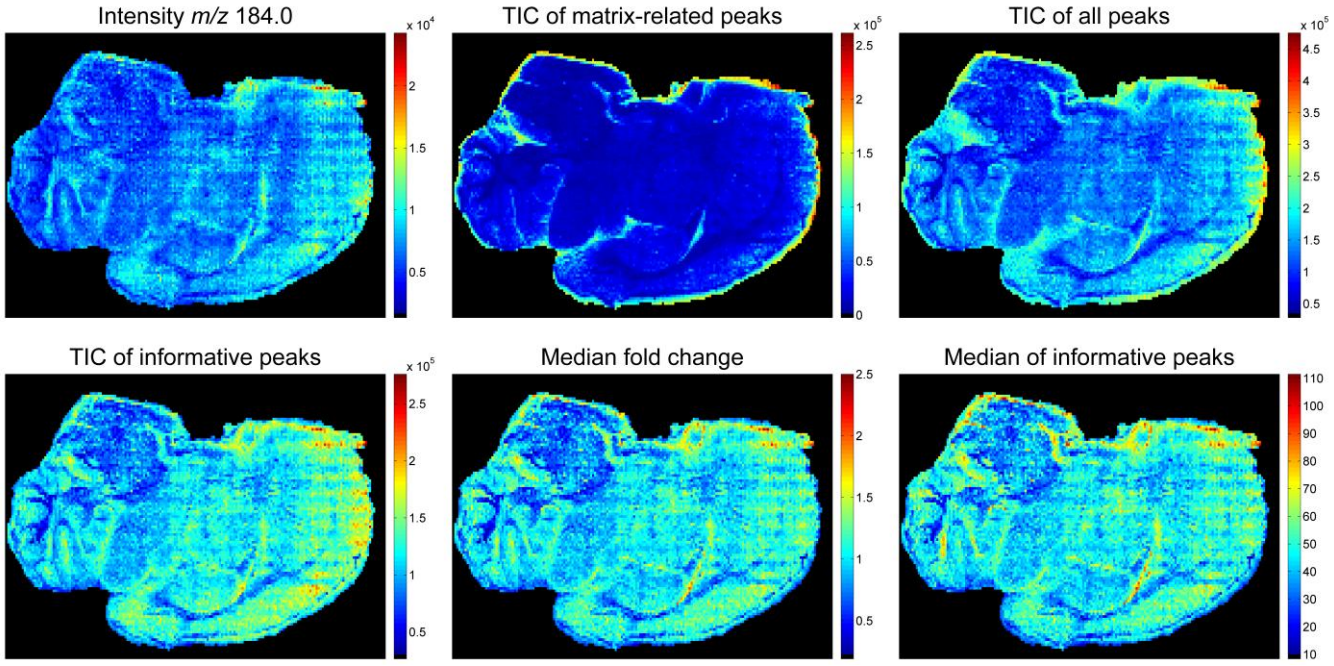


Figure 3. Normalization of MALDI MSI data of the sagittal rat brain section. The color scale for each image represents the factor by which the spectrum in an individual pixel would be divided for six normalization methods. Red represents the division by a higher factor and blue a lower factor.

1. Normalization to the intensity of a reference molecule is one method of normalizing the pixels.[12] Here, an *m/z* which is representative of a homogeneously distributed endogenous species is selected, such that normalization corrects for analytical and not biological variation. In this study, we normalized to *m/z* 184.0, which is the phosphocholine head group, and is expected to be relatively homogeneously distributed in this tissue type, which is reflected in the peak intensities that are displayed in *Figure 3*. Normalization to a reference molecule in MALDI remains a challenging task: variations in concentration, relative detection efficiency and adduct formation complicate the use of endogenous molecules.[36] This method would be more appropriate if an external, labeled calibration molecule was used as an internal standard,[16] but this is complicated by practical issues such as choice of compound and deposition method and may require extensive optimization.

2. Normalization to the TIC of "matrix-related peaks" is a correction for uneven matrix coating, which is performed by dividing peak intensities in a pixel by the sum of the peak intensities of all variables that were deleted with approach **A** for peak selection (the 3526 peaks that correlated with the intensities of the matrix ions). Normalization to matrix peaks is based on the assumption that more analyte is measured if there is more matrix signal present, as it is the matrix that co-crystallizes with the analytes. However, discrepancies in this relation are highlighted in *Figure S-2*, which shows that a false adjustment of the peak intensities for pixels could be the result of normalization to the TIC of matrix-related peaks.

These findings concur with published studies reporting that the degree of analyte incorporation in the matrix may vary, and that different analytes can be heterogeneously incorporated within the matrix crystals,[3, 31] which would cause this normalization method to be inappropriate.

3. Normalization to the TIC of all data involves the division of peak intensities in a pixel by the sum of all original 4750 variables, and is identical to normalization to the mean signal intensity (except

for a uniform scaling factor related to the number of variables). Although this is the most commonly used normalization method,[6-8, 10, 22, 25-26, 30] the problems observed for normalization to matrix-related peaks persist in this approach: high matrix concentration without co-crystallized analyte, crystal inhomogeneity and differential ionization efficiency will negatively affect the accuracy of this normalization. Because of these reasons, normalization to the TIC of all 4750 variables can create a false difference between compared pixels, see *Figure S-2* and *Figure 5 C*. As with the normalization to matrix-related peaks, a "halo" effect is also observed for the TIC of all data, where a higher TIC is observed around the edge of the sample (see *Figure 3*).[30] This information should, in fact, be used during quality control.

4. **Normalization to the TIC of "informative peaks"** is the division of peak intensities in a pixel by the sum of the 564 peak intensities after variable selection (the normalization factor is shown in *Figure 3*). Here, the rationale is that biological constituents might vary in concentration across different regions, but these will largely average out. The normalization based on informative peaks seems robust to the matrix-related issues discussed above.

It should be noted that different biological complexity, e.g. across different tissues, could negatively skew normalization to the TIC of informative peaks, as well as normalization to the TIC of all peaks. TIC-based normalization methods are also prone to the influence of a few high-intensity peaks (see also *Figure 5 C*),[26] and low-quality (noisy) pixels are problematic.

5. **Probabilistic quotient normalization** (PQN) is based on the median fold change of all peak intensities with respect to a reference spectrum (most commonly the median of the analyzed data).[37] This robust method may be less appropriate for normalization of MSI data, as is illustrated in *Figure S-3*, where the distribution of fold changes (based on the 564 selected peaks) for six anatomical regions are shown based on their different spectral profiles. Although the grey matter areas of the hippocampus,

cerebral cortex and cerebellar cortex show roughly symmetrical, Gaussian fold change distributions, supporting the use of PQN, the corpus callosum, optic chiasm and cerebellar white matter show uneven distributions around the median fold change, with fold changes for e.g. the cerebellar white matter of up to 60 (see *Figure S-3 C*). This characteristic of different tissue composition, present in MSI data, is prohibitive for the application of PQN, which is based on the assumption that less than 50% of the variables change across the different spectra.

6. Normalization by histogram matching is the process of converting the measured peak intensities to the values of the median spectrum, with respect to the rank order. Thus, the *m/z* variable with the highest intensity in a pixel is replaced with the highest value of the median spectrum, the second highest *m/z* variable is replaced with the second highest value in the median spectrum, etc.[35] This is a very robust approach, based on rank orders rather than intensity values, but the disadvantage is that the potentially informative intensity values are discarded. Moreover, the different composition of tissues could result in different peak intensity distribution histograms, which would reduce the effectiveness of this normalization method.

7. Normalization to the median intensity of the informative peaks is the division of peak intensities in a pixel by the median of the 564 peak intensities after variable selection (the normalization factor is shown in *Figure 3*). This is a compromise approach given the advantages and disadvantages of the methods listed so far: it does not assume constant molecular composition (as does PQN) or peak intensity distribution (as does histogram matching), it is not affected by differential co-crystallization (because only the selected peaks are included) and it is more robust to a few high-intensity peaks (unlike TIC-based normalizations, see also the comparison of normalization results for the simulated data set in *Figure 5 C*). Although the median intensity of the selected peaks is only an approximation to the normalization factor, and no theoretical basis can be given to further support this pragmatic

normalization method, it seems to be robust and is intended to be able to deal with most data sets.

Normalization should make the ratio of non-differentially abundant molecular features equal to one, and this was assessed by evaluating box-plots showing the distribution of fold changes of the peak intensities with respect to the median spectra for three different anatomical regions, as shown in *Figure 4*. After normalization to the median intensity of informative peaks, the fold changes are centered around one, indicating that this normalization method results in roughly similar spectral peak intensity distributions within each of the three anatomical regions (similar plots for normalization to TIC of informative peaks and histogram matching are shown in *Figure S-4*).

It should be noted that normalization to the median of informative peaks does not require comparable spectra for the different pixels, as PQN does, and thus the highly different spectral profiles of e.g. the optic chiasm and the hippocampus could now be more easily compared, with a reduced normalization bias. Erroneous spectra, e.g. without any signal or with an unusually low signal-to-noise ratio, would be inflated by most of the above mentioned methods (normalization to a reference compound could cope with the artifact in certain cases), and these spectra should regardless of normalization method be identified and removed prior to further data interpretation.

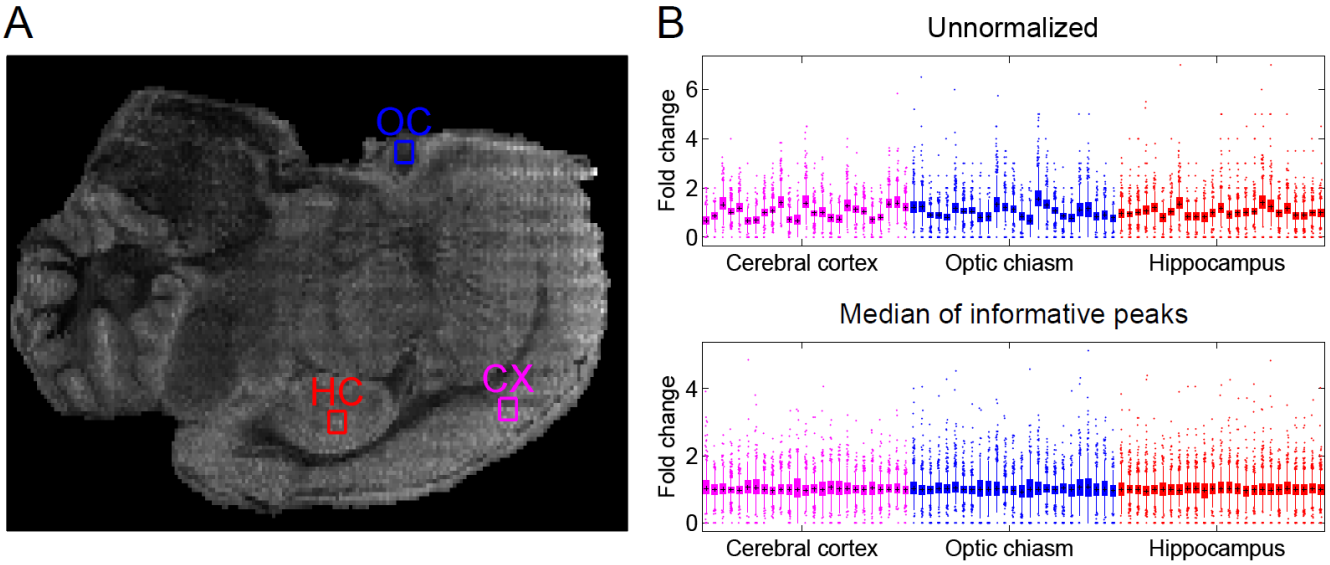


Figure 4. Effect of normalization to the median of informative peaks. (A) Three anatomical regions of 25 pixels were selected. Legend: OC: optic chiasm; HC: hippocampus; CX: cerebral cortex. (B) Box-plots show the distribution of the fold changes with respect to the median of each anatomical region (before and after normalization); *m/z* variables for which the median was zero in any of the regions were excluded.

Data transformation

For this MSI data set, the variance increases with peak intensity (see *Figure S-5*), as was confirmed by calculating the correlation between the mean and standard deviation of the peak intensity of individual peaks evaluated across all pixels. Logarithmic scaling was performed (see *Supporting Information*) to remove this heteroscedasticity, and increase the importance of lower intensity but structurally informative variables in subsequent modeling steps,[27] which is especially appropriate if a peak selection is performed, such as in this study. After transformation, mean-centering of the data was performed, which subtracts the average intensity for each *m/z* variable.

Effects of normalization on single *m/z* images and multivariate data analysis

The processing of the MSI data, namely peak and pixel selection, normalization, log transformation and mean-centering, resulted in a reduced data set of 13176 pixels × 564 *m/z* variables. *Figure 5* A shows images for five *m/z* values based on unnormalized and normalized data (normalized to the reference peak *m/z* 184.0, TIC of all peaks and the median of informative peaks); see *Figure S-6* for results of the other normalization methods. Unnormalized images show artifacts associated with laser oversampling, manifested as stripes in the ion image, these anomalies are clearly the result of analytical variation and not a genuine biological feature. Normalization to the TIC of all peaks does not fully compensate for these irregularities, while normalization to a selected reference peak appears to introduce speckled noise to the images. These ion intensities reveal less tissue anatomy and distinctness between regions than the newly suggested robust normalization approach to the median of informative peaks.

The robustness of normalization to the median of informative peaks is especially clear for the simulated data set with extreme peak intensities, shown in *Figure 5 C*: 3 regions of 5 by 5 pixels had their signal at *m/z* 772.4 artificially increased. The result of extreme peak intensity values in localized regions on the TIC-based normalization is clear from *Figure 5 C*: the acute peak intensity reductions in

the specific regions for TIC-based normalization are reflected as odd green squares in the hippocampus. In contrast, the normalization to the median intensity of selected peaks is not at all affected by a highly prevalent but localized peak (comparable to the insulin-rich regions in the pancreas as highlighted by Deininger *et al.*[26]). Thus TIC or other sum- or mean-based approach are prone to highly incorrect normalization factors for extreme high-intensity, localized signals, and the normalization based on median signal intensity is an elegant method to overcome this challenge and avoid detrimental normalization artifacts.

To evaluate the effect of the different normalization approaches on multivariate modeling, principal component analysis (PCA) was performed on the various data sets. PCA models the main variance in the data, and each pixel has a new set of coordinates in the calculated model, the scores, based on the *m/z* peak intensities for that pixel. The scores on a given principal component (PC) in the PCA model are used to color-code the image: pixels with high scores for a given PC will be shown in red, and pixels with low scores for this PC are shown in blue. The scores on PC 1 are biologically relevant for all data sets, irrespective of normalization, and show a clear distinction between grey and white matter (see *Figure S-7*). From *Figure 5 B*, it is clear that PC 2 is used to model analytical variation related to edge effects (TIC of all peaks), raster effects (horizontal stripes in the unnormalized data, especially visible in the cerebral cortex), and non-anatomical intensity variations for normalization to the single ion intensity at *m/z* 184.0. In contrast, anatomically relevant variation is being modeled on PC 2 for the data set that was normalized to the median of informative peaks, whilst these other data sets model a similar variation in PC 3. PCA results were comparable for the data normalized with histogram matching, the median or TIC of the informative peaks, and the probabilistic quotient normalization.

It should be emphasized that the two most commonly used normalization methods, i.e. normalization to the TIC of all peaks or a reference peak, gave markedly different PCA results compared to the normalization method here suggested, which is the result of PC 2 being used for the modeling of uncorrected analytical variation. It should also be noted that from these analyses it was especially clear

that normalization is necessary, as unnormalized data showed a characteristic raster effect.

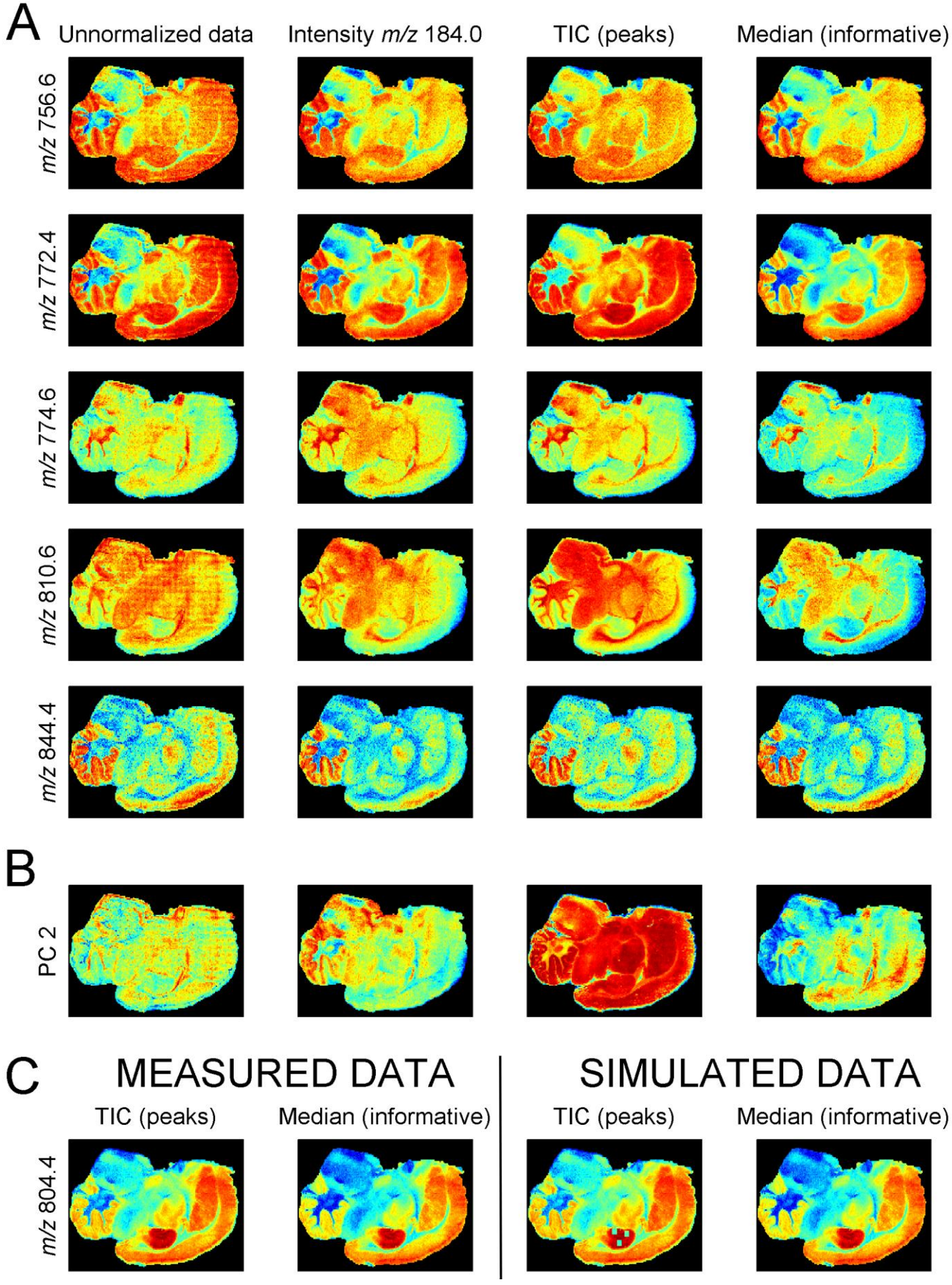
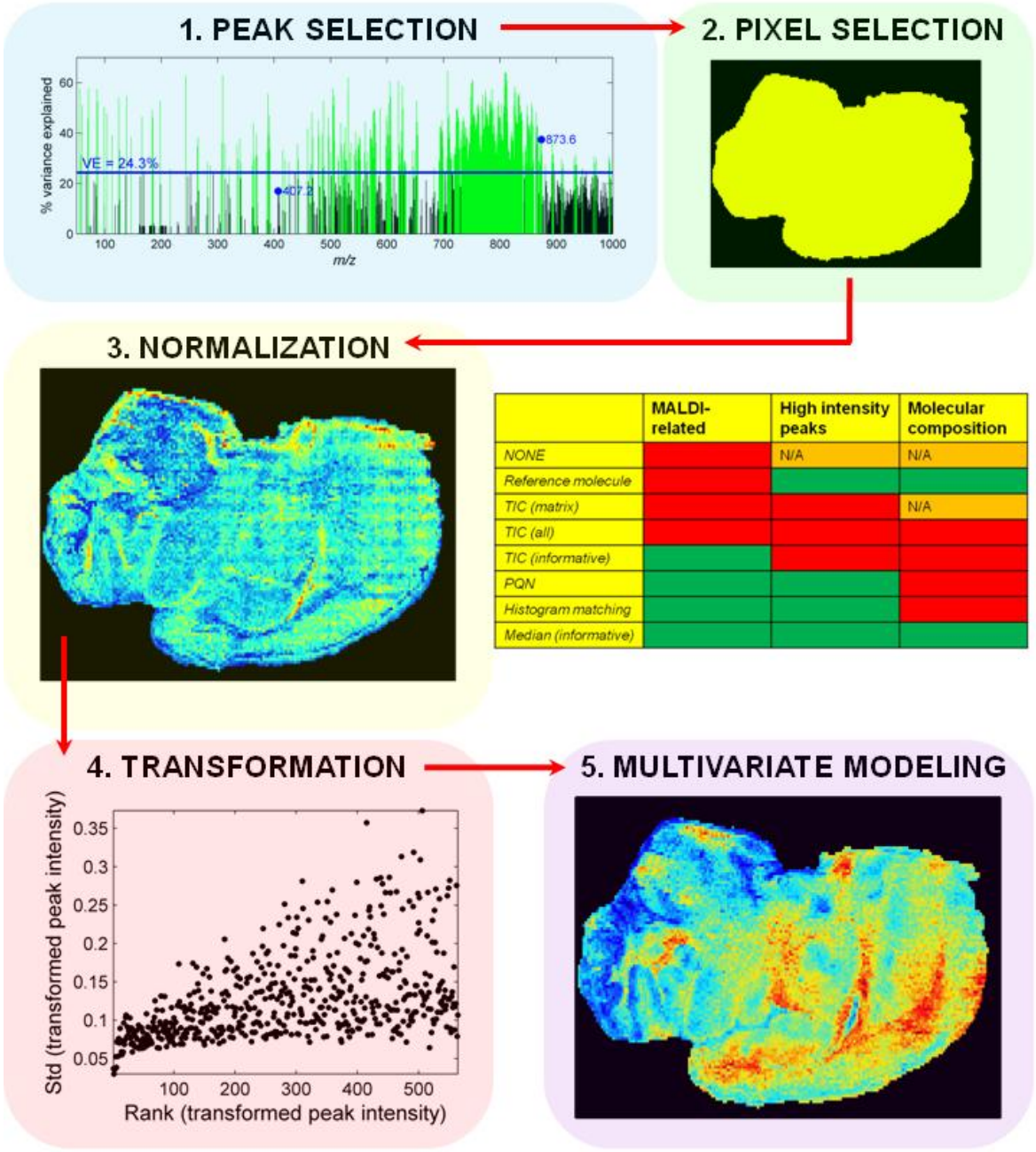

A
Unnormalized data
Intensity m/z 184.0
TIC (peaks)
Median (informative)
m/z 756.6
m/z 772.4
m/z 774.6
m/z 810.6
m/z 844.4
B
PC 2
C
MEASURED DATA
SIMULATED DATA
TIC (peaks)
Median (informative)
TIC (peaks)
Median (informative)
m/z 804.4

Figure 5. Results of normalization on single *m/z* images and principal component analysis. (A) MALDI images for five selected *m/z* values are color-coded with intensity (after log-transformation, blue for the lowest signal, red for high) and show the changes incurred by different normalization methods. (B) The score value of each pixel on the second principal component (PC 2) in a principal component analysis was used for color-coding. Legend: TIC (peaks): normalization to the TIC of all peaks; Median (informative): normalization to the median of informative peaks. The results for other normalization methods and other principal components are shown in *Figure S-6* and *Figure S-7*. (C) Image for *m/z* 804.4 in the original data set, for normalization the TIC of all peaks and the median of informative peaks, and the results of these two normalization methods for *m/z* 804.4 after the signal at *m/z* 772.4 was artificially increased in 3 regions (5 by 5 pixels) in the hippocampus.

Summary of data processing stream:

It is difficult to conduct a systematic evaluation of all possible processing approaches, because there are many combinatorial options of various methods for each different processing stage, and more importantly because there is no objective evaluation criterion: there is no clear metric that could be used to assess which of the different processing options is better. Therefore, the pragmatic processing choices used here to process the MALDI MSI data were based on rational arguments, such as pareto-efficiency for the *VE*-threshold as a variable selection cut-off, and a discussion of the assumptions and robustness for the different normalization methods has been provided. A flowchart for the processing approach suggested in this paper is shown in *Figure 6*. Note that changing the order for any of these steps can have a large effect on the resulting data set, and it is necessary to identify informative peaks in order to perform the proposed pixel selection and normalization procedures.

Data processing approaches should be robust and transferable to other data sets. A choice of necessary processing steps and a tailoring of parameters may be required for different *m/z* ranges and acquisition methods. Nonetheless, preliminary results on other sample types, including fresh (unfixed) brain tissue, demonstrated that the data processing workflow in *Figure 6* was directly applicable for various data sets.



	MALDI-related	High intensity peaks	Molecular composition
NONE		N/A	N/A
Reference molecule			
TIC (matrix)			N/A
TIC (all)			
TIC (informative)			
PQN			
Histogram matching			
Median (informative)			

Figure 6. Flowchart of MSI data processing. The consecutive processing steps for MSI data, including peak and pixel selection, normalization and data transformation, followed by multivariate modeling, are shown. A table indicating the properties of the different normalization methods is added in step 3 as a simplified summary, for more details please consult the text and the *Supporting Information*.

CONCLUSIONS

Data processing is a critical, non-trivial step for reliable information recovery from spatially resolved molecular profiles obtained by MALDI MS imaging, and this work suggests some solutions for a number of current bottlenecks in the processing of MALDI MSI data. The assumptions of various normalization methods are often not fulfilled, for example due to effects of the MALDI process, causing normalization involving matrix peaks to be ineffective, and the non-standard composition of different tissues. An alternative normalization based on the median intensity of informative peaks is suggested and should be more robust to these commonly occurring problems; although this method is unlikely to fully correct all possible experimentally and analytically induced differences, it is clearly an improvement compared to the commonly used normalization methods. The proposed work-flow, with simple and pragmatic processing steps, is not theoretically limited to any type of tissue. Intelligent handling and review of MSI data is expected to become increasingly important in the future, as this method increases in popularity and more applications are realized. We present here a first step towards a more sophisticated and accessible data analysis of these vast, highly informative, data sets. We hope this will assist in increasing the opportunities for MALDI MSI in existing and new applications, and to increase confidence in the reliability and reproducibility of the data, further establishing MSI as an invaluable and complementary tool to standard histological approaches.

ACKNOWLEDGMENTS

This work was supported by the award of an RSC PhD studentship to JMF and an EPSRC/RSC studentship to CLC. We thank Dr. Kirill Veselkov for useful discussions. The BRC is acknowledged for financial support to JMF and OC.

SUPPORTING INFORMATION

Supporting Information available as noted in the text. This material is available free of charge via the Internet at http://pubs.acs.org.

REFERENCES

1. Stoeckli, M.; Chaurand, P.; Hallahan, D. E.; Caprioli, R. M. *Nat. Med.* **2001**, *7*, 493-496.
2. Chaurand, P.; Caprioli, R. M. *Electrophoresis* **2002**, *23*, 3125-3135.
3. Chughtai, K.; Heeren, R. M. A. *Chem. Rev.* **2010**, *110*, 3237-3277.
4. McDonnell, L. A.; Heeren, R. M. A. *Mass Spectrom. Rev.* **2007**, *26*, 606-643.
5. Reyzer, M. L.; Caprioli, R. M. *Curr. Opin. Chem. Biol.* **2007**, *11*, 29-35.
6. Kang, S.; Shim, H. S.; Lee, J. S.; Kim, D. S.; Kim, H. Y.; Hong, S. H.; Kim, P. S.; Yoon, J. H.; Cho, N. H. *J. Proteome Res.* **2010**, *9*, 1157-1164.
7. Deininger, S. O.; Ebert, M. P.; Futterer, A.; Gerhard, M.; Rocken, C. *J. Proteome Res.* **2008**, *7*, 5230-5236.
8. Rauser, S.; Marquardt, C.; Balluff, B.; Deininger, S. O.; Albers, C.; Belau, E.; Hartmer, R.; Suckau, D.; Specht, K.; Ebert, M. P.; Schmitt, M.; Aubele, M.; Hofler, H.; Walch, A. *J. Proteome Res.* **2010**, *9*, 1854-1863.
9. Murphy, R. C.; Hankin, J. A.; Barkley, R. M. *J. Lipid Res.* **2009**, *50*, S317-S322.
10. Sugiura, Y.; Konishi, Y.; Zaima, N.; Kajihara, S.; Nakanishi, H.; Taguchi, R.; Setou, M. *J. Lipid Res.* **2009**, *50*, 1776-1788.
11. Benabdellah, F.; Touboul, D.; Brunelle, A.; Laprevote, O. *Anal. Chem.* **2009**, *81*, 5557-5560.
12. Burrell, M. M.; Earnshaw, C. J.; Clench, M. R. *J. Exp. Bot.* **2007**, *58*, 757-763.
13. Bunch, J.; Clench, M. R.; Richards, D. S. *Rapid Commun. Mass Spectrom.* **2004**, *18*, 3051-3060.
14. Khatib-Shahidi, S.; Andersson, M.; Herman, J. L.; Gillespie, T. A.; Caprioli, R. M. *Anal. Chem.* **2006**, *78*, 6448-6456.
15. Seeley, E. H.; Caprioli, R. M. *Proc. Natl. Acad. Sci. USA* **2008**, *105*, 18126-18131.
16. Rohner, T. C.; Staab, D.; Stoeckli, M. *Mech. Ageing Dev.* **2005**, *126*, 177-185.
17. Yang, Y. L.; Xu, Y. Q.; Straight, P.; Dorrestein, P. C. *Nat. Chem. Biol.* **2009**, *5*, 885-887.
18. Agar, N. Y. R.; Malcolm, J. G.; Mohan, V.; Yang, H. W.; Johnson, M. D.; Tannenbaum, A.;

Agar, J. N.; Blacks, P. M. *Anal. Chem.* **2010**, *82*, 2621-2625.

19. Stoeckli, M.; Staab, D.; Staufenbiel, M.; Wiederhold, K. H.; Signor, L. *Anal. Biochem.* **2002**, *311*, 33-39.

20. Ketterlinus, R.; Hsieh, S.-Y.; Teng, S. H.; Lee, H.; Pusch, W. *Biotechniques* **2005**, *38*, S37-S40.

21. Clerens, S.; Ceuppens, R.; Arckens, L. *Rapid Commun. Mass Spectrom.* **2006**, *20*, 3061-3066.

22. Djidja, M. C.; Claude, E.; Snel, M. F.; Francese, S.; Scriven, P.; Carolan, V.; Clench, M. R. *Anal. Bioanal. Chem.* **2010**, *397*, 587-601.

23. Dill, A. L.; Ifa, D. R.; Manicke, N. E.; Costa, A. B.; Ramos-Vara, J. A.; Knapp, D. W.; Cooks, R. G. *Anal. Chem.* **2009**, *81*, 8758-8764.

24. Norris, J. L.; Cornett, D. S.; Mobley, J. A.; Andersson, M.; Seeley, E. H.; Chaurand, P.; Caprioli, R. M. *Int. J. Mass spectrom.* **2007**, *260*, 212-221.

25. Hanselmann, M.; Kothe, U.; Kirchner, M.; Renard, B. Y.; Amstalden, E. R.; Glunde, K.; Heeren, R. M. A.; Hamprecht, F. A. *J. Proteome Res.* **2009**, *8*, 3558-3567.

26. Deininger, S. O.; Cornett, D. S.; Paape, R.; Becker, M.; Pineau, C.; Rauser, S.; Walch, A.; Wolski, E. *Anal. Bioanal. Chem.* **2011**, *401*, 167-181.

27. Wagner, M. S.; Graham, D. J.; Ratner, B. D.; Castner, D. G. *Surf. Sci.* **2004**, *570*, 78-97.

28. Carter, C. L.; McLeod, C. W.; Bunch, J. *J. Am. Soc. Mass. Spectrom.* **2011**, *22*, 1991-1998.

29. McDonnell, L. A.; van Remoortere, A.; de Velde, N.; van Zeijl, R. J. M.; Deelder, A. M. *J. Am. Soc. Mass. Spectrom.* **2010**, *21*, 1969-1978.

30. McDonnell, L. A.; van Remoortere, A.; van Zeijl, R. J. M.; Deelder, A. M. *J. Proteome Res.* **2008**, *7*, 3619-3627.

31. Dai, Y. Q.; Whittal, R. M.; Li, L. *Anal. Chem.* **1996**, *68*, 2494-2500.

32. van den Berg, R. A.; Hoefsloot, H. C. J.; Westerhuis, J. A.; Smilde, A. K.; van der Werf, M. J. *BMC Genomics* **2006**, *7*, 142.

33. Borgaonkar, S. P.; Hocker, H.; Shin, H. J.; Markey, M. K. *Omics* **2010**, *14*, 115-126.

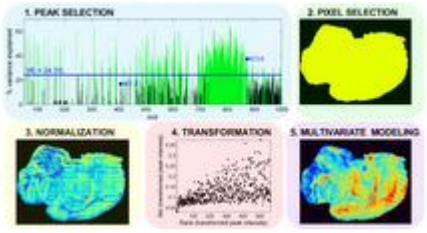
34. Meuleman, W.; Engwegen, J. Y. M. N.; Gast, M. C. W.; Beijnen, J. H.; Reinders, M. J. T.; Wessels, L. F. A. *BMC Bioinformatics* **2008**, *9*, 88.

35. Veselkov, K. A.; Vingara, L. K.; Masson, P.; Robinette, S. L.; Want, E.; Li, J. V.; Barton, R. H.; Boursier-Neyret, C.; Walther, B.; Ebbels, T. M.; Pelczer, I.; Holmes, E.; Lindon, J. C.; Nicholson, J. K. *Anal. Chem.* **2011**, *83*, 5864–5872.

36. Hankin, J. A.; Murphy, R. C. *Anal. Chem.* **2010**, *82*, 8476-8484.

37. Dieterle, F.; Ross, A.; Schlotterbeck, G.; Senn, H. *Anal. Chem.* **2006**, *78*, 4281-4290.

For TOC only


1. PEAK SELECTION
2. PIXEL SELECTION
3. NORMALIZATION
4. TRANSFORMATION
5. MULTIVARIATE MODELING