# Imaging Mass Spectrometry Data Reduction: Automated Feature Identification and Extraction

Liam A. McDonnell,[a] Alexandra van Remoortere,[a] Nico de Velde,[b] René J. M. van Zeijl,[a] and André M. Deelder[a]

[a] Biomolecular Mass Spectrometry Unit, Department of Parasitology, Leiden University Medical Center, Leiden, The Netherlands
[b] Hogeschool Leiden, Leiden, The Netherlands

Imaging MS now enables the parallel analysis of hundreds of biomolecules, spanning multiple molecular classes, which allows tissues to be described by their molecular content and distribution. When combined with advanced data analysis routines, tissues can be analyzed and classified based solely on their molecular content. Such molecular histology techniques have been used to distinguish regions with differential molecular signatures that could not be distinguished using established histologic tools. However, its potential to provide an independent, complementary analysis of clinical tissues has been limited by the very large file sizes and large number of discrete variables associated with imaging MS experiments. Here we demonstrate data reduction tools, based on automated feature identification and extraction, for peptide, protein, and lipid imaging MS, using multiple imaging MS technologies, that reduce data loads and the number of variables by >100×, and that highlight highly-localized features that can be missed using standard data analysis strategies. It is then demonstrated how these capabilities enable multivariate analysis on large imaging MS datasets spanning multiple tissues. (J Am Soc Mass Spectrom 2010, 21, 1969–1978) © 2010 American Society for Mass Spectrometry

Matrix assisted laser desorption/ionization mass spectrometry (MALDI MS) [1] can generate biomolecular profiles that describe the levels of hundreds of distinct biomolecules directly from tissue [2]. Spatially-correlated analysis, MALDI imaging MS, can simultaneously reveal how each of these biomolecules varies in heterogeneous tissue samples [3, 4].

There is growing evidence that imaging MS is having an impact in disease detection, particularly cancer [5]. The differential profiles found in tumors have been used to identify specific peptides and proteins that could act as biomarkers. Following the histologic annotation of a tissue section by a pathologist, the profiles obtained from histologically distinct regions are compared. The comparison of large numbers of tissues enables the specificity of these biomarkers to be ascertained. Such histology-defined analyses have been used to identify candidate biomarkers from prostate cancer [6, 7], lung cancer [8], and ovarian cancer [9].

One of the potential advantages of MALDI imaging MS is that it can define the regions of a tissue based on their biomolecular signatures and thereby identify those regions displaying signatures associated with a tumor but which have not yet undergone morphological transformation or regions that are not morphologically distinct using established histopathologic tools. MALDI imaging MS has found colorectal carcinoma related proteins in histologically benign polyps [10], revealed proteins specific to tumor interface zones and how normal tissue adjacent to the tumor expresses many of the molecular characteristics of the tumor [11]. This untargeted molecular analysis has even provided evidence for intra-tumor molecular heterogeneity [12].

The ability to detect changes in disease-associated protein expression independent of histologic annotation has several potential clinical applications:

1. Identification of sub-regions within tumors (intra-tumor heterogeneity within well defined histological subgroups).
2. Identification of regions indicating biomolecular change before morphological transformation.
3. Differentiation of tumors with overlapping histologies (within well defined histological subgroups).

Key to the development of such histology-independent analyses are tools for efficiently analyzing the very large datasets generated by imaging MS. An imaging MS analysis creates a mass spectrum for every pixel of the image, each of which contains a multitude of peaks that describe the mass and intensity of specific biomolecules. A variety of data analysis tools have been used to investigate these rich spatio-chemical datasets [12–

Address reprint requests to Dr. L. McDonnell, Biomolecular Mass Spectrometry Unit, Department of Parasitology, Leiden University Medical Center, Albinusdreef 2, 2333ZA Leiden, The Netherlands. E-mail: l.a.mcdonnell@lumc.nl.


1044-0305/10/$32.00
doi:10.1016/j.jasms.2010.08.008

Published online August 21, 2010
Received April 29, 2010
Revised August 10, 2010
Accepted August 13, 2010

18], however these molecular histology investigations have involved a very small numbers of tissues owing to the very large data loads produced by MALDI imaging MS experiments.

The number of discrete channels in a MALDI imaging MS experiment can also hinder the application of many data analysis tools. For example the number of floating point operations of a widely used singular value decomposition algorithm, as used in principal component an analysis, is given by

$$flops = 14 \cdot k \cdot N^2 + 8 \cdot N^3 \quad (1)$$

where $k$ is the number of pixels and $N$ the number of variables, or discrete mass channels, in a mass spectrum [19]. The number of channels, 60 k in a TOF instrument measured with full resolution or 500 k for an FTICR, needs to be reduced to efficiently analyze the intrinsic spatial variation in the tissue's biomolecular content.

Table 1 reports the data load and number of discrete mass channels that can be expected from an imaging MS analysis of a small clinical tissue array containing twenty large tissue samples (2 $cm^2$) and imaged with a spatial resolution of 100 $\mu$m. The experiment would yield 400 k individual mass spectra. Depending on how the mass spectrometry is performed the total data load can be 100–400 Gb, and the number of channels per mass spectrum 50 k–500 k, much too large for simultaneous analysis of the entire dataset (a 32 bit computer has a maximum memory allocation of 4 Gb). However, if the datasets were reduced to the mass spectral features detected in each experiment the reduced data load would allow molecular histology of small tissue arrays (Table 1).

Several methods for reducing this data load have been investigated. Broersen et al. [20] analyzed multiple tissues by selecting a reduced mass range and reducing the mass resolution by a factor of 1000 [20]. However this strategy necessarily compromises mass resolution to a degree that might be deemed unacceptable, and could lead to the merging of mass spectral peaks. The smoothing and data-reduction capabilities of discrete wavelet transformation have been used in MALDI mass spectrometry for improved quantitation [21]. van de Plas et al. have reported a preliminary investigation of the use of wavelet transforms for MALDI imaging MS [22]. A 1D wavelet transform of each pixel's mass spectrum yielded an ~8× reduction in data load, and number of channels, but retained almost all spatio-chemical information. Note: This degree of reduction is probably an underestimate, reflecting the very low mass resolution of the original dataset, just 6490 samples spanning the *m/z* range 2800–25,000. Three-dimensional wavelet based data reduction exceeding 100× has been reported for hyperspectral imaging from focal plane array detectors [23]. When using wavelet-based data reduction techniques, any subsequent analysis is performed on the wavelet coefficients and not the measured masses. An inverse discrete wavelet transform (IDWT) can be used to reconstruct the *m/z* signals [24] or the regions

**Table 1.** Typical dataloads for clinical tissue analyses using MALDI imaging mass spectrometry prior to feature extraction and following feature extraction. Original dataloads were obtained from experiments using typical experimental parameters for peptide imaging (MALDI-reflectron TOF), protein imaging (MALDI-linear TOF) and lipid imaging (MALDI-FTICR)

| Data load in imaging mass spectrometry | | | |
|---|---|---|---|
| Number of tissues | 20 | | |
| Size of tissue | 2 $cm^2$ | | |
| Spatial resolution | 100 $\mu$m | | |
| Number of pixels per tissue = | 20000 | | |
| Total number of pixels = | 400000 | | |
| Number of mass spectra = | 400000 | | |
| | Time-of-flight | | FTICR |
| | Peptides | Proteins | Lipids |
| Mass range | 1000–5500 | 3.5 k–30 k | 400–1500* |
| Digitization rate | 1 GHz | 1 GHz | 714 kHz* |
| Number of channels per spectrum | 65 k | 100 k | 500 k* |
| Data load per mass spectrum | 250 | 400 | 1000 Kb |
| Total data load per tissue | 5 | 8 | 20 Gb |
| Total data load for 20 tissues | 100 | 160 | 400 Gb |
| Number of biomolecules detected = | 100 | 150 | 500 |
| Total number of values = | 4.00E+07 | 6.00E+07 | 2.00E+08 |
| Data load per value (64 bit) | 8 | 8 | 8 Bytes |
| Total data load (64 bit) | 305 | 458 | 1526 Mb |
| Total data load (32 bit) | 153 | 229 | 763 Mb |
| Data reduction factor (64 bit) | 328 | 350 | 262 |
| Data reduction factor (32 bit) | 655 | 699 | 524 |

*The digitization rate of the FTICR is determined by the lower mass of the selected mass range (and therefore the highest cyclotron frequency [31]). In this case *m/z* 400 corresponds to a cyclotron frequency of approximately 357 kHz and so the FTICR transient is sampled at 714 kHz (Nyquist criterion [32]). These settings, using a 1 M datapoint FID ensured the full lipid mass range is measured with a mass resolution exceeding 100 k on a 9.4T FTICR.

identified in the analysis of the wavelet-compressed data can be used to extract region-of-interest spectra from the original MS dataset. For pathology applications, mass spectral accuracy and specificity are crucial to be able to identify the molecules highlighted by an analysis and to subsequently validate any results using independent tests, e.g., immunohistochemistry for a specific protein [6].

We have previously distilled protein imaging MS datasets to the mass spectral features-of-interest by extracting a manually defined peak list [17] from each pixel's mass spectrum. However, such manual annotation is incompatible with the high throughput imaging capabilities [25] necessary for analyzing clinical cohorts and lacks objectivity. Here we report the development of entirely automated feature detection and extraction algorithms for imaging MS datasets that encompasses improved feature detection, applicability to multiple MS platforms, and can be used for the automated reduction of the datasets from entire sample cohorts.

## Experimental

Tumor tissue samples obtained from surgical resection or post-mortem specimens were snap frozen in liquid isopentane and then stored at −80 °C until sectioning. 5 μm thick tissue sections were cut at −20 °C using a cryomicrotome and stained with hematoxylin and eosin (H an E) to check diagnosis and viability of the tissue. For the MALDI imaging mass spectrometry experiments, 12 μm thick tissue sections were cut at −20 °C and thaw mounted onto conductive glass slides (Delta Technologies, Stillwater, MN, USA). The tissues were then slowly brought to room temperature in a dessiccator and prepared for MALDI analysis of the tissue's peptides, proteins or lipids. All tissue samples were handled in a coded fashion, according to Dutch national ethical guidelines (code for proper secondary use of human tissue, Dutch Federation of Medical Scientific Societies).

### *Peptide Imaging*

A uniform coating of α-cyano-4-hydroxycinnamic acid (HCCA) was added using an ImagePrep device (Bruker Daltonics, Bremen, Germany) and a solution of 10 mg/mg HCCA in 70:30 AcN: 0.1% TFA(aq.). MALDI imaging MS was then performed using an Ultraflex III MALDI-TOF/TOF (Bruker Daltonics), 100 μm pixel size, and 800 laser shots per pixel (50 laser shots per position of a random walk within each pixel). Data acquisition, preprocessing (smoothing and baseline subtraction of each pixel's MALDI mass spectrum), and data visualization/process verification were performed using the Flex software suite (FlexControl 3.0, FlexAnalysis 3.0, FlexImaging 2.1).

### *Protein Imaging*

The tissues were washed in isopropanol and sinapinic acid (SA) matrix was added using an ImagePrep device and a solution of 20 mg/mL SA in 70% isopropanol: 0.1% TFA(aq.). MALDI Imaging MS experiments were then performed using an Autoflex III MALDI-TOF (Bruker Daltonics), 100 μm pixel size, and 600 laser shots per pixel (50 laser shots per position of a random walk within each pixel). Data acquisition, preprocessing (baseline subtraction of each pixel's MALDI mass spectrum), and data visualization/process verification were performed using the Flex software suite (FlexControl 3.0, FlexAnalysis 3.0, FlexImaging 2.1).

### *Lipid Imaging*

A uniform coating of 2,5-dihydroxyacetophenone matrix was added using a 20 μm stainless steel sieve [26] and the tissue analyzed using either an UltrafleXtreme MALDI-TOF/TOF or a 9.4T APEX-Ultra MALDI FTICR (both Bruker Daltonics). Experiments performed with the UltrafleXtreme mass spectrometer used a 100 μm pixel size with 500 laser shots per pixel (100 laser shots per position of a random walk within each pixel) and were acquired using the Flex software suite (FlexControl 3.3, FlexImaging 2.1, FlexAnalysis 3.3). Experiments performed on the APEX-Ultra FTICR mass spectrometer used a 200 μm pixel size with 450 laser shots per pixel (50 laser shots per position of a random walk within each pixel) and were acquired in fully automated mode using FlexImaging 2.1, Hystar 3.4, ApexControl 3.0. Lipid peak assignments were made by comparing each peak's accurate mass measurement with the LIPID MAPS database (http://lipidmaps.org, mass accuracy ±0.005 Da).

Data reduction and data analysis was performed using custom scripts written in Matlab (ver. 7.4.0. Mathworks). A full description of the automated feature identification and extraction algorithm is included as Supplementary Information, which can be found in the electronic version of this article. In brief, the algorithm first calculates several different mass spectral representations of the dataset, the mean mass spectrum, the basepeak mass spectrum and their TIC normalized analogues. The formulae used for the mean and basepeak representations are

$$\overset{\text{mean}}{\bar{I}_m = \frac{\sum_{1}^{j} I_{m,j}}{j}} \quad \overset{\text{basepeak}}{I_{m,\max} = \max_{1}^{j}(I_{m,j})} \tag{2}$$

where $I_m$ is the intensity of the $m^{th}$ channel in the mass spectrum and $j$ is the pixel number. It will be shown that the use of multiple mass spectral representations enables established peak picking methods to be used to identify spatio-chemical features more effectively than current methods.

Automated feature detection is then performed on each mass spectral representation using an adapted

form of the LIMPIC program [27]. This feature detection method has been developed for the reliable detection and quantitation of protein peaks in MALDI-TOF profiles but is also adept at identifying the sharper peaks produced by the higher resolution reflectron-TOF and FTICR mass spectrometers. The algorithm uses a set of statistical tests to decompose the mass spectrum into signal, baseline and noise.

Firstly the spectrum is divided into a series of *m/z* blocks and the kurtosis of each block is calculated to identify blocks that do not contain significant peaks (kurtosis $<1$). For the TOF spectra the average values of the blocks free of significant peaks were used to interpolate the baseline drift, which was then subtracted from the spectrum. The high dynamic range and high mass resolution of the FTICR measurements, experimental mass resolution of $\approx$120k at *m/z* 760, led to spectra characterized by a very low baseline. It was found that the slow baseline drift of the FTICR spectra was most easily estimated by interpolating between the minima of all blocks. An *m/z*-dependent estimate of the noise in each mass spectral representation was then performed by calculating the standard deviations of the signals in the blocks free of significant peaks and interpolating through the entire mass range. The peak picking algorithm then searches for localized maxima within the mass spectra defined by the minimum peak width. Only those peaks exceeding a user-defined signal-to-noise threshold and a percentage basepeak threshold (percentage of most intense peak in spectrum) are retained. The use of the four mass spectral representations enabled conservative estimates of the S/N thresholds to be used to distinguish the peaks that provided high quality images from noisy images (Supplemental Figure 5 and Supplemental Figure 6).

The peak lists from each mass spectral representation are combined into a final, collated peak list, which is then used to extract cross sections of the imaging MS dataset at these peak values. These feature detection and extraction algorithms have been incorporated into a single workflow that can be applied to multiple MS methods (linear TOF, reflectron TOF, and FTICR), and which enables multiple experiments to be selected for data reduction without additional user intervention.

The results reported in this manuscript describe the further development of our previous data extraction method based on the manual definition of protein peaks-of-interest [17]. The new algorithm includes automated and improved feature detection, automated feature extraction, for multiple MS methodologies, and molecular classes, which can be applied for the automated reduction of imaging MS datasets from entire sample cohorts.

## Results and Discussion

### *Example Datasets*

Figure 1 shows an example MALDI imaging MS dataset and illustrates the peptide spectra and images that can be obtained from a human pancreas tissue section. MALDI imaging MS datasets are normally analyzed by using the mean mass spectrum as the mass-spectral user-interface of the dataset. Each of the peptide peaks can be selected to display that peptide's image. The images obtained from the human pancreas tissue include distributions covering most of the tissue, as well as highly localized features corresponding to the endocrine peptides produced by the pancreas' islets of Langerhans, for example amylin (the tissue used in this experiment was from an elderly patient with type II diabetes).

Closer inspection of the pancreas MALDI imaging MS dataset revealed that the mean spectrum, shown in black, could under-represent highly localized peptides. Figure 1 also shows the mass spectrum obtained from a specific islet of Langerhans. This spectrum contains high intensity peaks that were not apparent in the mean spectrum of the tissue, and whose images displayed a high degree of localization.

Most data analyses, and especially automated analyses, are based on the calculation of the tissue's mean spectrum. The residual background from the surrounding tissue can overwhelm signals from highly localized peptides and proteins, even after background subtraction of each pixel's mass spectrum (as was performed here, see Supplementary Information for details). Thus, one of the potential advantages of MALDI imaging MS, namely its ability to analyze heterogeneous tissues may be undermined by the omission of localized features in the tissue.

Biomarker discovery experiments based on a comparative analysis of histologically defined regions can detect localized differences if the histologic annotation is highly detailed. However, the annotation typically spans larger regions containing multiple cell types. Furthermore, as biomarker discovery experiments attempt to identify candidate biomarkers with high sensitivity and specificity, the analysis purposely avoids the intra-heterogeneity and inter-heterogeneity of the tissues. Such histology-directed analyses are unable to distinguish between tissues that are not set apart before the analysis, for example tumors having overlapping histologies; nor can they detect the molecular changes that occur before histologic transformation or occur in areas bordering tumor masses [10, 28].

As explained in the introduction imaging MS has been used to identify regions with differential molecular signatures that could not be distinguished using established histologic tools. The potential of so-called molecular histology to provide an independent, complementary analysis of clinical tissues lies in its ability to image the distributions of hundreds of biomolecules from the same tissue section, without a priori knowledge of the tissue and spanning multiple molecular classes. The large datasets generated by imaging MS have limited molecular histology experiments to studies encompassing a very small number of tissues. This is especially true for lipid and metabolite imaging MS,

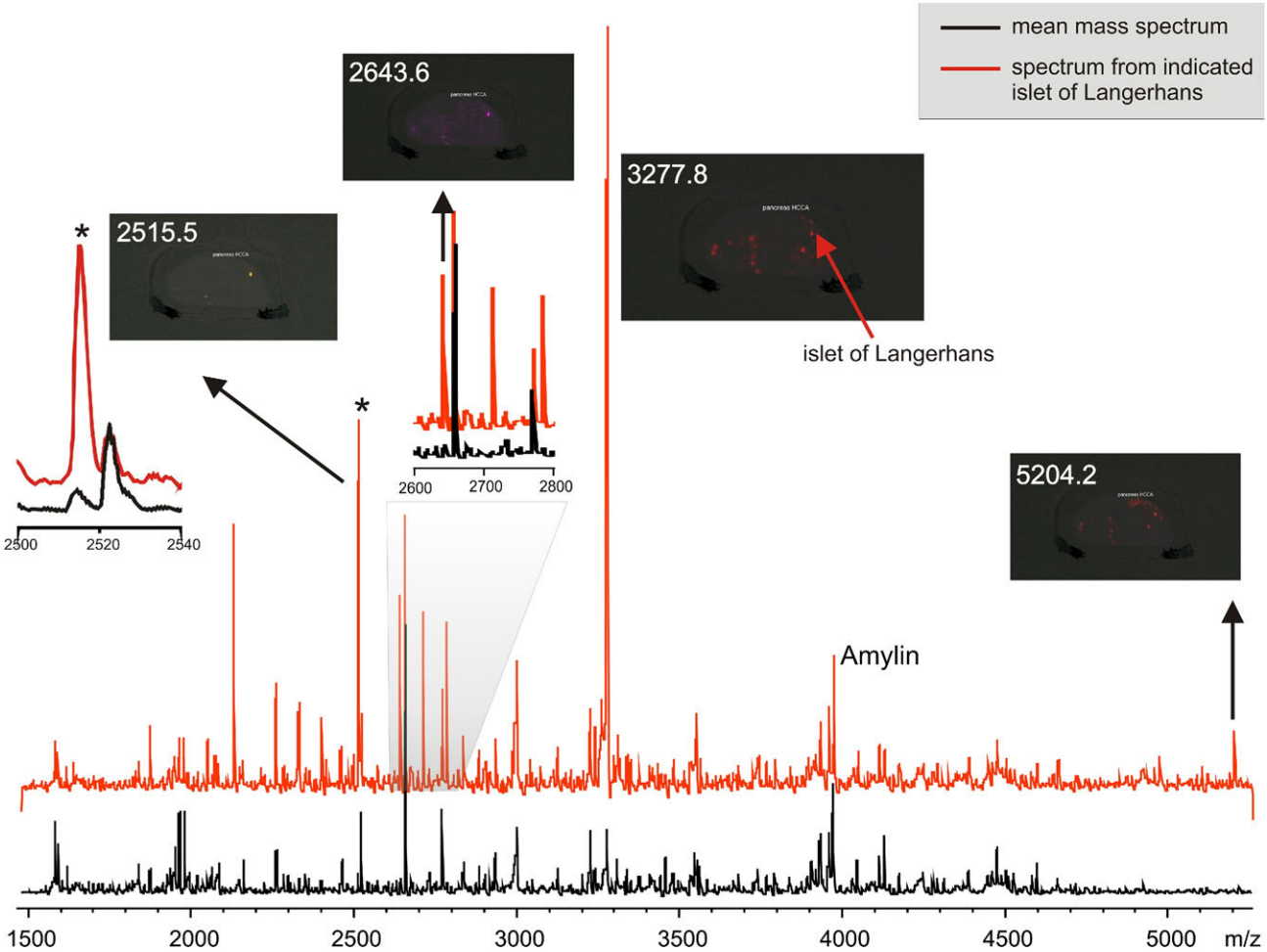


**Figure 1.** Close examination of a MALDI-TOF imaging MS analysis of peptides in human pancreas tissue reveals highly localized peptides, which are not visible in the dataset's mean mass spectrum.

which are becoming the domain of high mass resolution/ion-mobility-mass spectrometry imaging because of the need to distinguish between ions of identical nominal mass. Figure 2 shows examples of MALDI imaging MS of lipids in human muscle and human brain tissues using ultra high mass resolution Fourier transform ion cyclotron resonance mass spectrometry (FTICR MS). The high mass resolution is necessary to

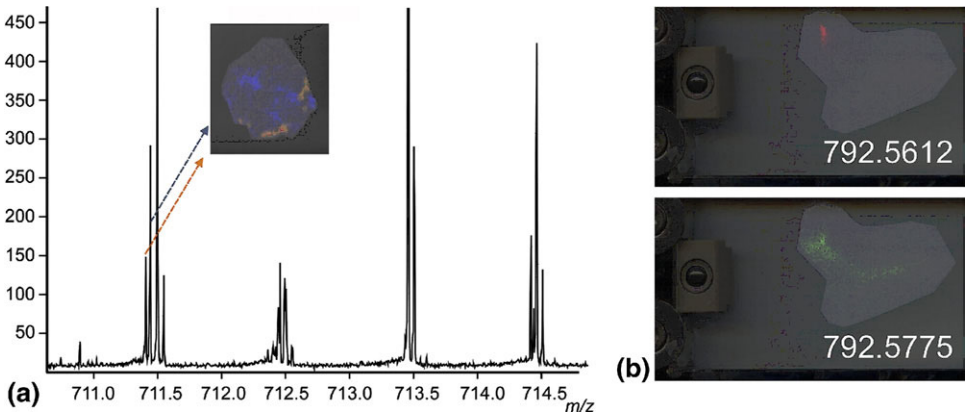


**Figure 2.** MALDI FTICR imaging MS analyses of lipids in human muscle (**a**) and human brain (**b**) tissues requires ultra high mass resolution to resolve the high number of isobaric peaks, which can have entirely different spatial distributions. Such high mass resolution experiments leads to very large file sizes: human muscle sample 32.3 Gb, human brain sample 62.6 Gb.

distinguish between the different lipid ions, which despite having very similar masses can have very different spatial distributions and be highly localized. The price of high mass resolution is large file sizes (Table 1).

## *Efficient and Effective Data Reduction*

Data reduction routines have been developed that import all of the required instrumental parameters and then read each pixel's mass spectrum from the proprietary data format. The raw or processed data can be imported from the MALDI-TOF instruments and the transients are imported from the FTICR, which are then zero-filled, apodized, Fourier transformed, and converted to the *m/z* domain. A set of mass spectral representations are then calculated to distinguish all features in the imaging MS dataset; this includes the dataset's mean spectrum, basepeak spectrum and their TIC normalized analogues. The basepeak spectrum displays the maximum intensity detected in the entire imaging dataset for every *m/z*, consequently higher intensity peaks are explicitly included irrespective of how localized they are in the tissue. Note: the basepeak representation will include the maximum 'background noise' value for every mass value; consequently it is crucial that every pixel's mass spectrum is processed using an effective smoothing and baseline subtraction routine before data reduction [17, 29]. A detailed description of the algorithm is provided as Supplementary Information.

Each mass spectral representation is the sum of many thousands of individual spectra. To aid feature detection each mass spectral representation is smoothed and baseline subtracted before the peak selection algorithm is applied. The peaks detected in each spectrum are then collated into a final peak list, which is subsequently used to extract each peak's intensity from every pixel and create the reduced image cube dataset. Figure 3 shows the four mass spectral representations. The blue line shows the unsmoothed mass spectral representations, the red line the smoothed and baseline subtracted mass spectral representations, and the green diamonds indicate the peaks contained in the final, collated peak list. Note: to illustrate which peaks were detected in each mass spectral representation, the green diamonds are plotted at the intensities of the detected peaks; if the peak was not detected, the intensity remains zero.

Using a S/N threshold of 4 and a local estimation of the spectrum's noise [27] the automated feature identi-

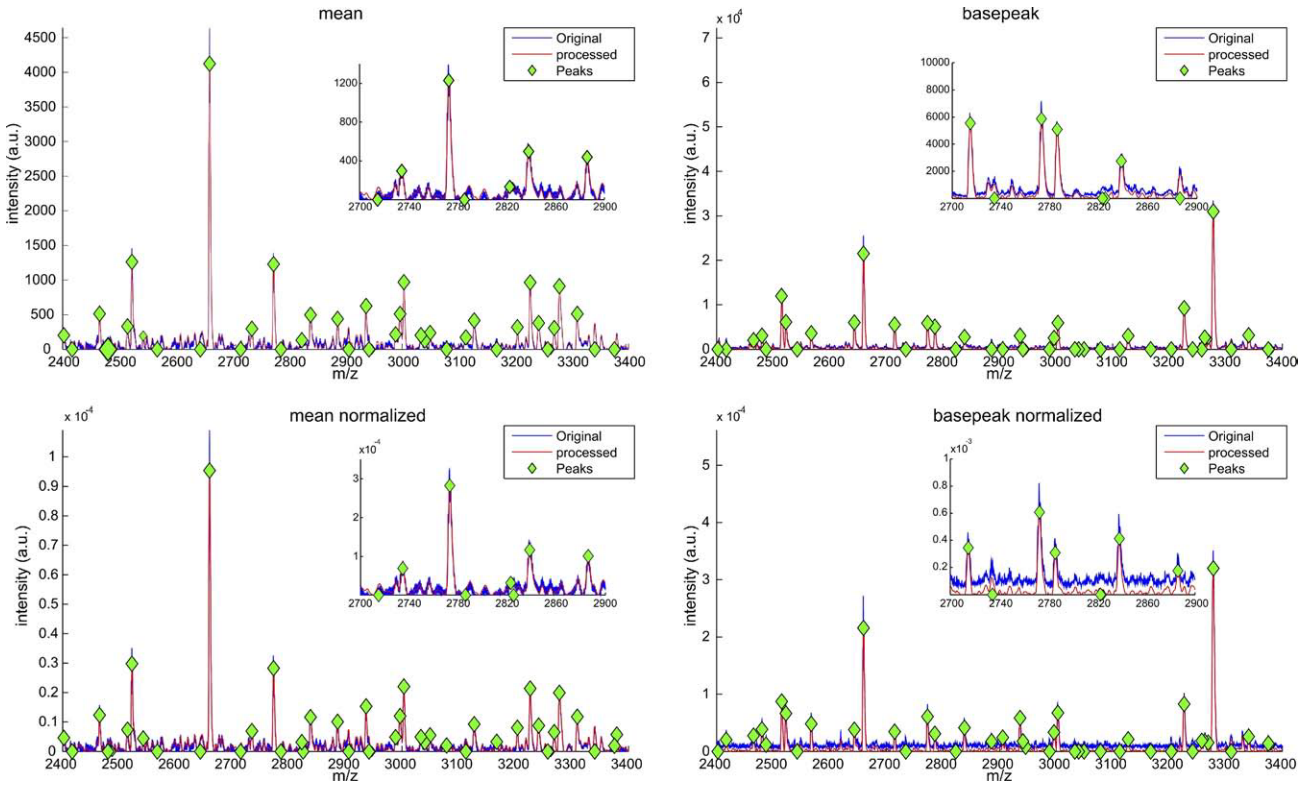


**Figure 3.** Automated feature detection of the peptide imaging MS dataset of human pancreas tissue shown in Figure 1. Four spectral representations of the dataset are calculated to ensure all peaks are detected: the mean spectrum displays the mean intensity of each mass across the entire dataset, the basepeak spectrum displays the maximum intensity of each mass across the entire dataset, and their TIC-normalized analogues ensure peaks under-represented due to inhomogeneous matrix coverage are also identified. The peaks from all four mass spectral representations are collated into a final peak list, which is then used to reduce the imaging MS dataset into an image cube.

fication and extraction algorithm detected 146 peptides in the pancreas MALDI imaging MS dataset, of which only 95 were detected in the mean spectrum, and reduced the data-load from 1.1 Gb for the complete dataset to just 3.4 Mb for the reduced (extracted) dataset. Figure 4 shows that peaks detected in the basepeak spectrum but not the mean spectrum can correspond to biomolecules with a highly localized distribution. In this case nine additional highly localized peptides were detected (see Supplementary Figure 9, for spectra and images of all nine peptides), and even for the peptides that were detected in the mean spectrum their relative magnification in the basepeak mass spectral representation indicated they may be localized.

An explicit quality control mechanism has been developed to assess the performance of the data reduction algorithm. The reduced data is exported in the proprietary peak list format to verify that the results of the data reduction reproduce the original imaging MS data recorded using FlexImaging (Supplementary Figure 8).

Figure 5 shows the results of the feature selection algorithm applied to the 32.3 Gb MALDI-FTICR imaging MS dataset of a human muscle biopsy previously shown in Figure 2, using conservative peak selection criteria (S/N threshold of 4 and a basepeak intensity threshold of 0.1%). The use of multiple mass spectral representations increased the number of peaks detected in the imaging MS dataset from 344 in the mean spectrum to 851. The subsequent removal of ions with a mass defect not consistent with lipids [30], which are due to species such as matrix clusters, led to a peak list containing 712 ions (Supplementary Figure 7). This peak list was then used to distill the entire 32.3 Gb MALDI-FTICR dataset to an image cube containing all 712 images. Using a 64-bit data format and retaining each pixel's TIC, the four mass spectral representations and a range of additional metadata (that may be used in subsequent data analyses), the reduced data load was 34.8 Mb. Figure 4 demonstrates that this data reduction scheme also identified lipids with localized distributions that were not selected in the mean mass spectral representation.

Further data reduction through summing the images of a specific lipid's isotopomers was investigated; 218 [M + 1] isotopomers and 21 [M + 2] isotopomers were detected within a mass accuracy of ±1 ppm. However, close inspection of the mean mass spectral representation revealed that many of the [M + 2] isotopomer peaks were only partially resolved. Evidently the 120 k mass resolution of the experiment was not sufficient to resolve all lipid ions obtained from the muscle tissue. The low statistical probability of lipid [M + 2] isomers means the potential contribution of [M + 2] isotopes will be small for all but the most intense lipids with a mass <1000 Da but not for larger lipids, such as

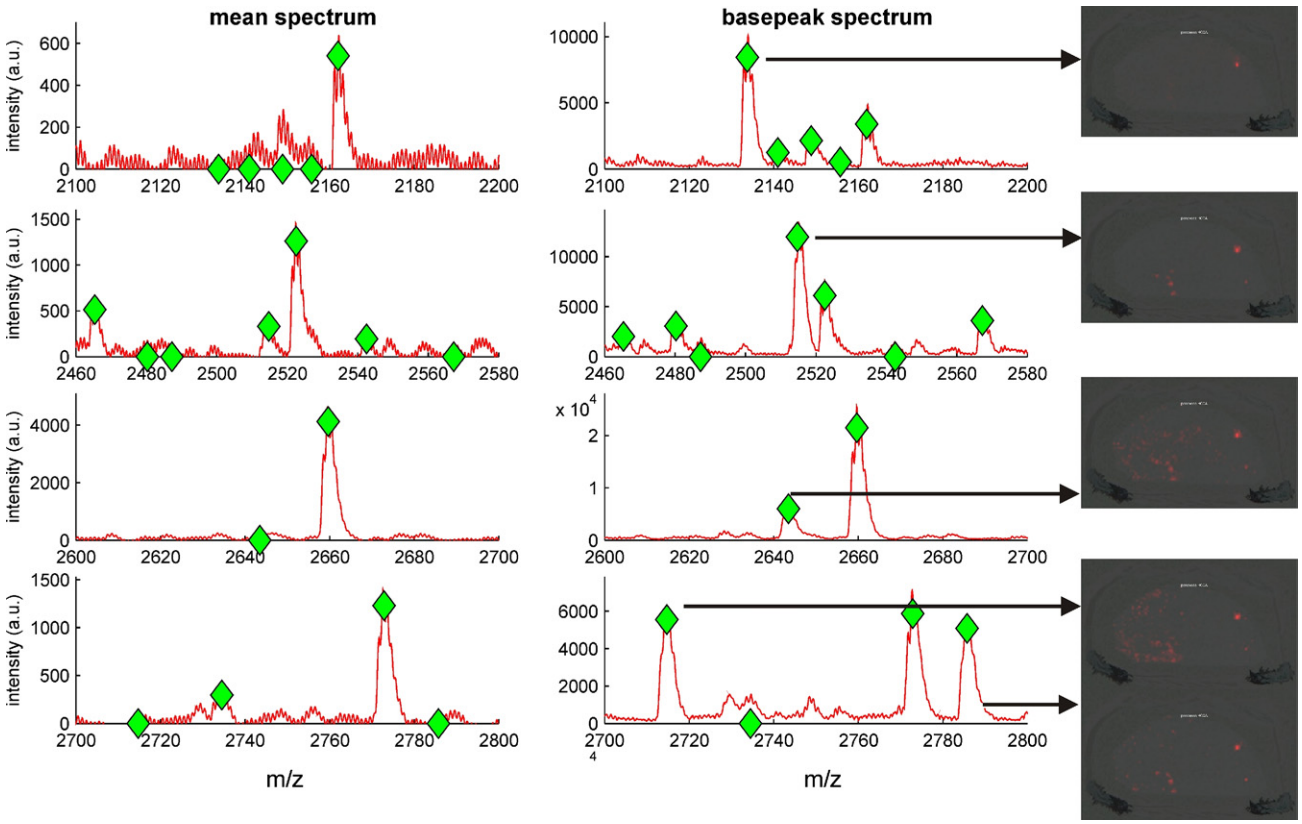


**Figure 4.** Peaks highlighted in the basepeak spectrum relative to the mean spectrum can indicate highly localized peptides.

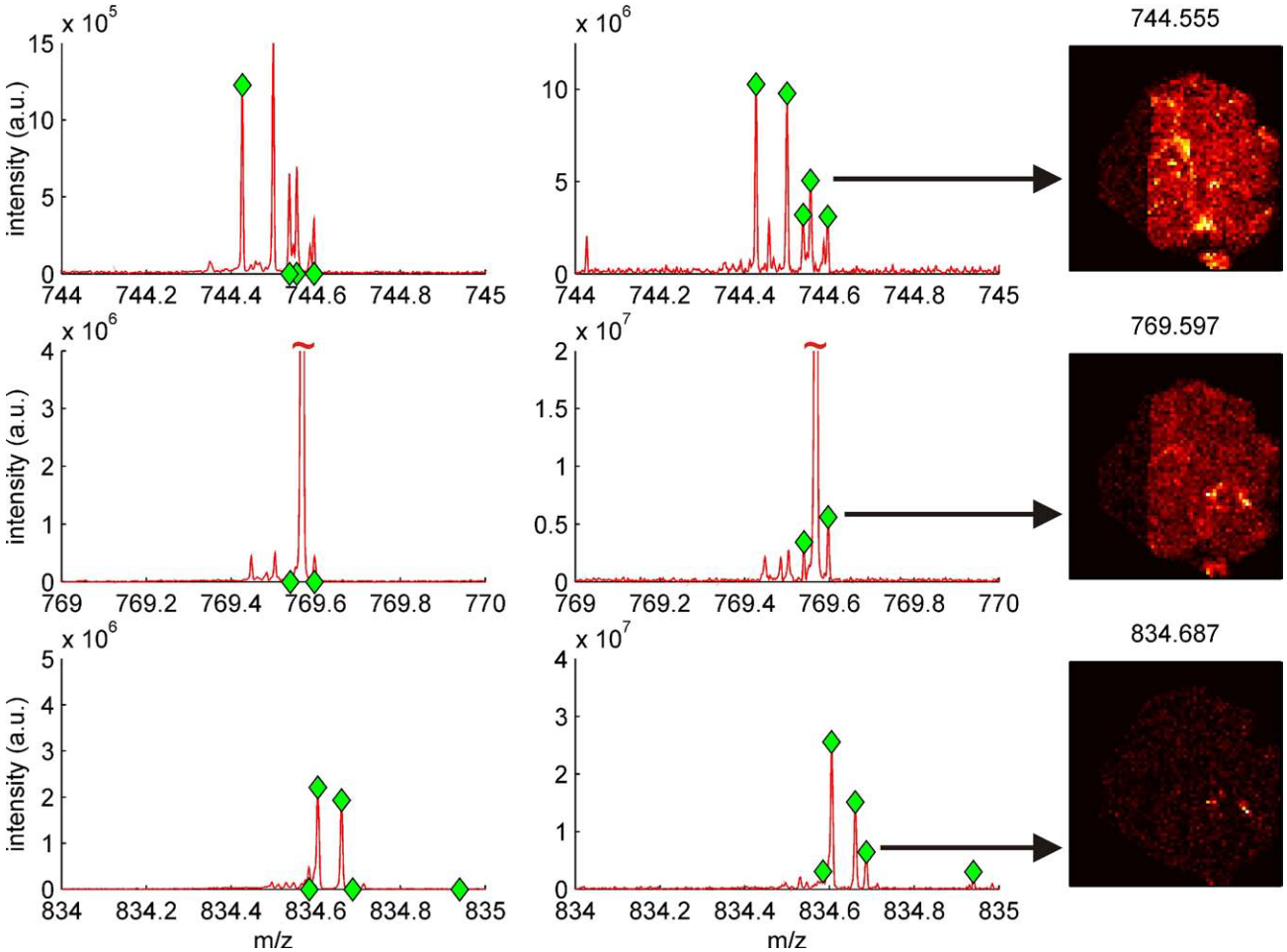


**Figure 5.** Automated data reduction of the MALDI FTICR human muscle dataset shown in Figure 2 identified 851 distinct peaks, 712 of which were lipids. Peaks highlighted in the basepeak spectrum relative to the mean spectrum displayed a high degree of localization.

gangliosides. Nevertheless, the ≈900× reduction in data load and number of variables provided by the algorithm is sufficient to enable simultaneous multivariate analysis of the data from multiple tissues analyzed using MALDI-FTICR-MS.

It may be argued that several of the localized peaks detected in the basepeak mass spectral representation would also have been detected in the mean mass spectral representation if the S/N = 4 and intensity threshold (0.1% of maximum) had been set much lower. However, such an approach would have led to the inclusion of a very large number of low intensity peaks that are associated with weak, highly pixilated and thus unreliable images (Supplementary Figure 5 and Supplementary Figure 6). The basepeak mass spectral representation allows highly localized distributions, described by high quality mass spectral peaks, to be distinguished from the weak, sporadic signals of unreliable images.

These data reduction routines efficiently reduce the large datasets generated by MALDI-TOF and MALDI-FTICR imaging MS experiments into an image cube containing the images of all features detected in the dataset, and even includes the highly localized species that are frequently missed using typical data analysis routines. Table 2 provides a summary of the performance of the data reduction routines for different MALDI imaging MS applications. Even when all data

**Table 2.** Performance of the data reduction algorithms on MALDI-TOF and MALDI-FTICR imaging MS datasets

| | Pancreas peptides | Muscle lipid |
|---|---|---|
| *Original data* | | |
| Number of pixels | 2155 | 3163 |
| Entire dataset | 1.1 Gb | 32.3 Gb |
| Mass spectrum size | 52476 channels | 1048576 channels |
| *Reduced data* | | |
| Number of images | 147 | 712 |
| Reduced dataset (64 bit) | 3.4 Mb | 34.8 Mb |
| *Reduction factor* | 330 | 928 |

The reduced datasets include all data that might be used during further analysis (each pixel's TIC for normalization, coordinates of each pixel, each mass spectral representation of the dataset, the experiment name, collated peak list, and peak intensities for each of the four mass spectral representations).

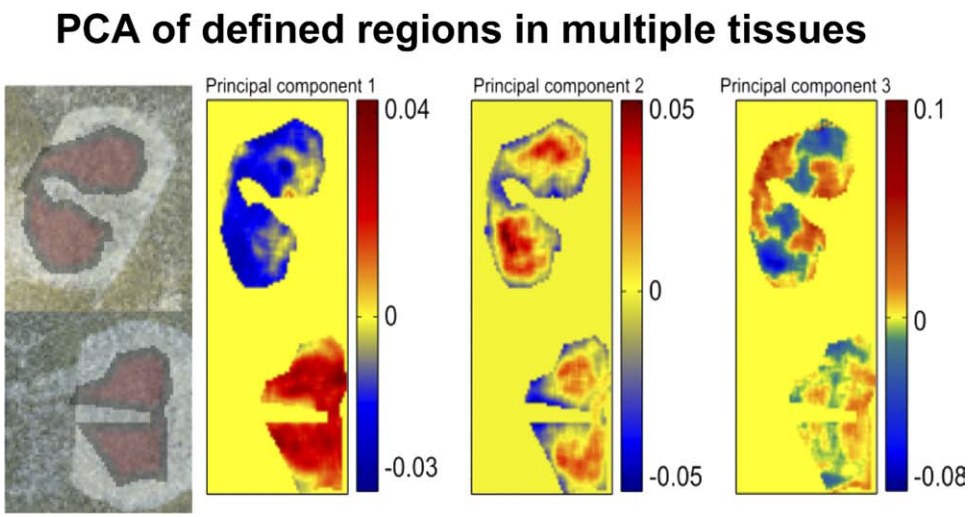


**Figure 6.** Principal component analysis of specific regions, defined by the distribution of a single protein, in two adjacent mouse brain tissue sections.

that might be used during further analysis is retained (each pixel's TIC for normalization, each pixel's coordinates, each mass spectral representation of the dataset, the experiment name, collated peak list, and collated peak intensities for each of the four mass spectral representations), the data loads associated with imaging MS datasets can be reduced by a factor of 300–900×.

The substantial reduction in data load and the number of discrete variables enable multiple datasets to be combined and their data simultaneously analyzed. Figure 6 shows an optical image of two mouse brain tissue sections that had been prepared for protein imaging MS and were subsequently analyzed using a MALDI TOF. After data reduction, an image mask corresponding to the distribution of a single protein (indicated in Figure 6) was applied and principal component analysis (PCA) performed to investigate the variance within these protein-specific regions. Figure 6 shows that the first principal component reveals the differences between the two tissues whereas the second and third principal components reveal the correlations within the defined regions.

The low mass resolution of peptide and protein peaks produced by MALDI and analyzed with a linear MALDI-TOF mass spectrometer can make automated peak selection less reliable. The low S/N of the peptide and protein peaks in each pixel's mass spectrum (each pixel analyzes a very small number of cells) exacerbates this problem; consequently on-the-fly detection of the peaks in each pixel's mass spectrum has been limited to Fourier transform-type instruments (the standard mode of operation of the Orbitrap). On-the-fly peak detection of each pixel's mass spectrum must balance the use of low peak detection thresholds (to select as many of the low S/N peaks as possible) with the increased amount of noise retained in the dataset. The use of the four mass spectral representations reported here, and the collation of the resulting peak lists, allows more demanding peak selection criteria to be used while also ensuring the detection of localized features.

## Conclusion

Data reduction algorithms have been developed for lipid, peptide, and protein imaging MS that effectively summarize each dataset by determining multiple mass spectral representations of each dataset and extracting the spatial variation (MS image) of each ion above a set of user-defined peak thresholds. The algorithms have been explicitly designed to include (and highlight) localized features that can be easily missed when using standard tools based on the dataset's mean mass spectrum, and provide a quality-control feature for testing the performance of the reduction using the original, proprietary, imaging MS analysis software. These data reduction tools begin to address the issue of the very large dataloads and number of variables in MALDI imaging MS. It is hoped that the research presented here will prompt a wider effort to establish best practice guidelines for data reduction in MALDI imaging MS, thus enabling the clinical potential of molecular histology to be fully investigated.

## Acknowledgments

The authors acknowledge funding for this work by the ZonMw Horizon program 'Multiplex Imaging of Tissue Arrays,' project number 93519026 (L.A.McD). Dr. S. M. Willems, Professor Dr. P. C. W. Hogendoorn, Dr. M. L. C. Maat-Schieman (all Leiden University Medical Center) and Professor R. A. H. Adan (University of Utrecht) are gratefully acknowledged for the tissue samples. L.A.McD. gratefully acknowledges Dr. Aswin Verhoeven for helpful discussions during the development of the data reduction algorithms.

## Appendix A
## Supplementary Material

Supplementary material associated with this article may be found in the online version at doi:10.1016/j.jasms.2010.08.008.

## References


1. Karas, M.; Bachmann, D.; Bahr, U.; Hillenkamp, F. Matrix-Assisted Ultraviolet Laser Desorption of Nonvolatile Compounds. *Int. J. Mass Spectrom. Ion Processes* **1987,** *78,* 53–68.
2. Chaurand, P.; Sanders, M. E.; Jensen, R. A.; Caprioli, R. M. Proteomics in Diagnostic Pathology—Profiling and Imaging Proteins Directly in Tissue Sections. *Am. J. Pathol.* **2004,** *165,* 1057–1068.
3. Cornett, D. S.; Reyzer, M. L.; Chaurand, P.; Caprioli, R. M. MALDI Imaging Mass Spectrometry: Molecular Snapshots of Biochemical Systems. *Nat. Methods* **2007,** *4,* 828–833.
4. McDonnell, L. A.; Heeren, R. M. A. Imaging Mass Spectrometry. *Mass Spectrom. Rev.* **2007,** *26,* 606–643.
5. McDonnell, L. A.; Willems, S. M.; Corthals, G. L.; van Remoortere, A.; van Zeijl, R. J. M.; Deelder, A. M. Imaging Mass Spectrometry in Cancer Research: Past Experiences and Future Possibilities. *J. Proteom.* **2010,** *73,* 1921–1944.
6. Cazares, L. H.; Troyer, D.; Mendrinos, S.; Lance, R. A.; Nyalwidhe, J. O.; Beydoun, H. A.; Clements, M. A.; Drake, R. R.; Semmes, O. J. Imaging Mass Spectrometry of a Specific Fragment of Mitogen-Activated Protein Kinase/Extracellular Signal-Regulated Kinase Kinase Kinase 2 Discriminates Cancer from Uninvolved Prostate Tissue. *Clin. Cancer Res.* **2009,** *15,* 5541–5551.
7. Schwamborn, K.; Krieg, R. C.; Reska, M.; Jakse, G.; Knuechel, R.; Wellmann, A. Identifying Prostate Carcinoma by MALDI-Imaging. *Int. J. Mol. Med.* **2007,** *20,* 155–159.
8. Groseclose, M. R.; Massion, P. P.; Chaurand, P.; Caprioli, R. M. High-Throughput Proteomic Analysis of Formalin-Fixed Paraffin-Embedded Tissue Microarrays Using MALDI Imaging Mass Spectrometry. *Proteomics* **2008,** *8,* 3715–3724.
9. Lemaire, R.; Menguellet, S. A.; Stauber, J.; Marchaudon, V.; Lucot, J. P.; Collinet, P.; Farine, M. O.; Vinatier, D.; Day, R.; Ducoroy, P.; Salzet, M.; Fournier, I. Specific MALDI Imaging and Profiling for Biomarker Hunting and Validation: Fragment of the 11S Proteasome Activator Complex, Reg $\alpha$ fragment, is a New Potential Ovary Cancer Biomarker. *J. Proteome Res.* **2007,** *6,* 4127–4134.
10. Pevsner, P. H.; Melamed, J.; Remsen, T.; Kogos, A.; Francois, F.; Kessler, P.; Stern, A.; Anand, S. Mass Spectrometry MALDI Imaging of Colon Cancer Biomarkers: A New Diagnostic Paradigm. *Biomarkers Med.* **2009,** *3,* 55–69.
11. Oppenheimer, S. R.; Mi, D.; Sanders, M. E.; Caprioli, R. M. Molecular Analysis of Tumor Margins by MALDI Mass Spectrometry in Renal Carcinoma. *J. Proteome Res.* **2010,** *9,* 2182–2190.
12. Deininger, S. O.; Ebert, M. P.; Fütterer, A.; Gerhard, M.; Röcken, C. MALDI Imaging Combined with Hierarchical Clustering as a New Tool for the Interpretation of Complex Human Cancers. *J. Proteome Res.* **2008,** *7,* 5230–5236.
13. Eijkel, G. B.; Kükrer Kaletaş, B.; van der Wiel, I. M.; Kros, M.; Luider, T. M.; Heeren, R. M. A. Correlating MALDI and SIMS Imaging Mass Spectrometric Datasets of Biological Tissue Surfaces. *Surf. Interface Anal.* **2009,** *41,* 675–685.
14. Hanselmann, M.; Kirchner, M.; Renard, B. Y.; Amstalden, E. R.; Glunde, K.; Heeren, R. M. A.; Hamprecht, F. A. Concise Representation of Mass Spectrometry Images by Probabilistic Latent Semantic Analysis. *Anal. Chem.* **2008,** *80,* 9649–9658.
15. Hanselmann, M.; Kothe, U.; Kirchner, M.; Renard, B. Y.; Amstalden, E. R.; Glunde, K.; Heeren, R. M. A.; Hamprecht, F. A. Toward Digital Staining using Imaging Mass Spectrometry and Random Forests. *J. Proteome Res.* **2009,** *8,* 3558–3567.
16. McCombie, G.; Staab, D.; Stoeckli, M.; Knochenmuss, R. Spatial and Spectral Correlations in MALDI Mass Spectrometry Images by Clustering and Multivariate Analysis. *Anal. Chem.* **2005,** *77,* 6118–6124.
17. McDonnell, L. A.; van Remoortere, A.; van Zeijl, R. J. M.; Deelder, A. M. Mass Spectrometry Image Correlation: Quantifying Co-Localization. *J. Proteome Res.* **2008,** *7,* 3619–3627.
18. van de Plas, R.; Ojeda, F.; Dewil, M.; van Den Bosch, L.; De Moor, B.; Waelkens, E. Prospective Exploration of Biochemical Tissue Composition Via Imaging Mass Spectrometry Guided by Principal Component Analysis. *Proceedings of the Pacific Symposium on Biocomputing 12 (PSB)*; Maui, Hawaii, January, 2007; pp 458–469.
19. Golub, G. H.; van Loanb, C. F. Matrix Computationsb. Baltimoreb: John Hopkins University Pressb, 1996b.
20. Broersen, A.; van Liere, R.; Altelaar, A. F. M.; Heeren, R. M. A.; McDonnell, L. A. Automated, Feature-Based Image Alignment for High-Resolution Imaging Mass Spectrometry of Large Biological Samples. *J. Am. Soc. Mass Spectrom.* **2008,** *19,* 823–832.
21. Coombes, K. R.; Tsavachidis, S.; Morris, J. S.; Baggerly, K. A.; Hung, M.-C., Kuerer, H. M. Improved Peak Detection and Quantification of Mass Spectrometry Data Acquired from Surface-Enhanced Laser Desorption and Ionization by Denoising Spectra with the Undecimated Discrete Wavelet Transform. *Proteomics* **2005,** *5,* 4107–4117.
22. van de Plas, R.; de Moor, B.; Waelkens, E. Discrete Wavelet Transform-based Multiivariate Exploration of Tissue via Imaging Mass Spectrometry. *Proceedings of the 23rd Annual ACM Symposium on Allied Computing (ACM SAC)*; Fortaleza, Brazil, March, 2008.
23. Vogt, F.; Banerji, S.; Booksh, K. Utilizing Three-Dimensional Wavelet Transforms for Accelerated Evaluation of Hyperspectral Image Cubes. *J. Chemometr.* **2004,** *18,* 350–362.
24. Bradley, A. P. *Proceedings of the 7th International Conference on Digital Image Computing: Techniques and Applications*; Sydney, Australia, December, 2003; p. 29–38.
25. McDonnell, L. A.; van Remoortere, A., van Zeijl, R. J. M., Dalebout, H.; Bladergroen, M. R.; André M. D. Automated Imaging MS: Toward High Throughput Imaging Mass Spectrometry. *J. Proteom.* **2009,** *73,* 1279–1282.
26. Puolitaival, S. M.; Burnum, K. E.; Cornett, D. S.; Caprioli, R. M. Solvent-Free Matrix Drycoating for MALDI Imaging of Phospholipids. *J. Am. Soc. Mass Spectrom.* **2008,** *19,* 882–886.
27. Mantini, D.; Petrucci, F.; Pieragostino, D.; Del Boccio, P.; Di Nicola, M.; Di Ilio, C.; Federici, G.; Sacchetta, P.; Comani, S.; Urbani, A. LIMPIC: A Computational Method for the Separation of Protein MALDI-TOF-MS Signals from Noise. *BMC Bioinformatics* **2007,** *8,* 101.
28. 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. Molecular Proteomics Imaging of Tumor Interfaces by Mass Spectrometry. *J. Proteome Res.* **2010,** *9,* 1157–1164.
29. Norris, J. L.; Cornett, D. S.; Mobley, J. A.; Andersson, M.; Seeley, E. H.; Chaurand, P.; Caprioli, R. M. Processing MALDI Mass Spectra to Improve Mass Spectral Direct Tissue Analysis. *Int. J. Mass Spectrom.* **2007,** *260,* 212–221.
30. Jones, J. J.; Stump, M. J.; Fleming, R. C.; Lay, J. O.; Wilkins, C. L. Strategies and Data Analysis Techniques for Lipid and Phospholipid Chemistry Elucidation by Intact Cell MALDI FTMS. *J. Am. Soc. Mass Spectrom.* **2004,** *15,* 1665–1674.
31. Marshall, A. G.; Hendrickson, C. L.; Jackson, G. S. Fourier Transform Ion Cyclotron Resonance Mass Spectrometry: A Primer. *Mass Spectrom. Rev.* **1998,** *17,* 1–35.
32. Kelly, P. C.; Horlick, G. Practical Considerations for Digitizing Analog Signals. *Anal. Chem.* **1973,** *45,* 518–527.