Chemo-informatic strategy for imaging mass spectrometry-based hyperspectral profiling of lipid signatures in colorectal cancer

Kirill A. Veselkov[a,1], Reza Mirnezami[b], Nicole Strittmatter[a], Robert D. Goldin[c], James Kinross[b], Abigail V. M. Speller[c], Tigran Abramov[d], Emrys A. Jones[a], Ara Darzi[b], Elaine Holmes[a], Jeremy K. Nicholson[a], and Zoltan Takats[a]

[a]Computational and Systems Medicine, Department of Surgery and Cancer, Faculty of Medicine, Imperial College London, London SW7 2AZ, United Kingdom; [b]Biosurgery and Surgical Technology, Department of Surgery and Cancer and [c]Centre for Pathology, Department of Medicine, Faculty of Medicine, St. Mary's Hospital, Imperial College London, London W2 1NY, United Kingdom; and [d]Department of Computer Science, Sevastopol National Technical University, Streletskaya Bay, Crimea 99053, Ukraine

Edited* by Burton H. Singer, University of Florida, Gainesville, FL, and approved December 9, 2013 (received for review June 7, 2013)

Mass spectrometry imaging (MSI) provides the opportunity to investigate tumor biology from an entirely novel biochemical perspective and could lead to the identification of a new pool of cancer biomarkers. Effective clinical translation of histology-driven MSI in systems oncology requires precise colocalization of morphological and biochemical features as well as advanced methods for data treatment and interrogation. Currently proposed MSI workflows are subject to several limitations, including nonoptimized raw data preprocessing, imprecise image coregistration, and limited pattern recognition capabilities. Here we outline a comprehensive strategy for histology-driven MSI, using desorption electrospray ionization that covers (*i*) optimized data preprocessing for improved information recovery; (*ii*) precise image coregistration; and (*iii*) efficient extraction of tissue-specific molecular ion signatures for enhanced biochemical distinction of different tissue types. The proposed workflow has been used to investigate region-specific lipid signatures in colorectal cancer tissue. Unique lipid patterns were observed using this approach according to tissue type, and a tissue recognition system using multivariate molecular ion patterns allowed highly accurate (>98%) identification of pixels according to morphology (cancer, healthy mucosa, smooth muscle, and microvasculature). This strategy offers unique insights into tumor microenvironmental biochemistry and should facilitate compilation of a large-scale tissue morphology-specific MSI spectral database with which to pursue next-generation, fully automated histological approaches.

Mass spectrometry imaging (MSI) of biological tissue sections can provide topographically localized biochemical information to supplement conventional histopathological classification systems (1–3). Together with emerging metabolomics-based profiling approaches, MSI represents a highly promising approach in molecular systems oncology (4, 5) and is increasingly being used for the discovery of next-generation cancer biomarker panels (6, 7). Among the MSI techniques currently available, the three most commonly used are matrix-assisted laser desorption ionization (MALDI) (2, 6), secondary ion mass spectrometry (SIMS) (8, 9), and desorption electrospray ionization (DESI) (10, 11). With each of these described approaches, operating characteristics and experimental parameters can be modulated to suit specific analytical objectives and can be customized for the identification of particular biomolecular species. Here, we have opted to use the DESI technique as there are several practical advantages with this method for metabolome-wide imaging studies, primarily attributable to lack of requirement for matrix deposition and ambient ionization, which requires minimal sample preparation (11, 12).

Currently MSI is likely to exert greatest influence at the prognostic and therapeutic stages of the disease continuum (Fig. 1), with three fundamental areas of application in cancer phenotyping. First, it offers a means of chemically mapping morphological regions of interest to develop next-generation prognostic and therapeutic biomarkers. Second, it permits compartmentalized assessment of the distribution and biochemical influence of chemotherapeutic agents and/or their downstream metabolites within different tissue regions, offering fresh insights into anticancer drug efficacy (13, 14). Third, MSI provides the opportunity to develop automated approaches for tissue classification based entirely on molecular ion patterns. Such automated, "machine-learned" strategies will lessen the logistical and financial burden being placed on pathology services in the modern cancer-screening era, while simultaneously ensuring quality control by minimizing interobserver variability (15).

Until now the routine clinical application of MSI approaches has been restricted by inherent time/cost demands and associated heavy analytical workload. However, recent advances in MS technology combined with the richness of generated molecular information should ensure the widespread adoption of MSI technologies in the near- to midterm. The major impediment to this progress currently centers on the choice of chemo-informatics workflow. The standard approach applied to MSI datasets involves a series of steps designed to reduce bioanalytical complexities for improved information recovery, followed by pattern recognition analysis and molecular pattern interpretation. Conventional workflows, integrated into software packages such as BioMap (Novartis), SpectViewer (CEA), DataCubeExplorer (AMOLF), and Mirion (JLU) or within commercial packages

Significance

Mass spectrometry imaging (MSI) technology represents a highly promising approach in cancer research. Here, we outline current roadblocks in translational MSI and introduce a comprehensive workflow designed to address current methodological limitations. An integrated bioinformatics platform is presented that allows intuitive histology-directed interrogation of MSI datasets. We show that this strategy permits the analysis of multivariate molecular signatures with direct correlation to morphological regions of interest, which can offer new insights into how different tumor microenvironmental populations interact with one another and generate novel region-of-interest specific biomarkers and therapeutic targets.

Author contributions: K.A.V., R.M., J.K., A.D., E.H., J.K.N., and Z.T. designed research; K.A.V., R.M., N.S., A.V.M.S., and Z.T. performed research; K.A.V., N.S., R.D.G., E.A.J., and Z.T. contributed new reagents/analytic tools; K.A.V., R.M., N.S., R.D.G., J.K., A.V.M.S., T.A., E.A.J., E.H., J.K.N., and Z.T. analyzed data; and K.A.V., R.M., N.S., R.D.G., J.K., A.V.M.S., T.A., E.A.J., A.D., E.H., J.K.N., and Z.T. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: kirill.veselkov04@imperial.ac.uk.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1310524111/-/DCSupplemental.

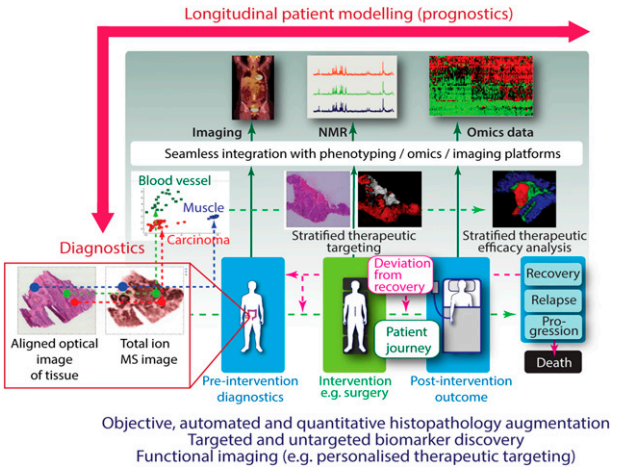


Fig. 1. MS-based imaging technology in clinical settings.

from instrument manufacturers such as Xcalibur (Thermo Fisher Scientific) and FlexImaging (Bruker Daltonics) have capabilities limited to basic preprocessing and browsing through selected ion images. There is currently strong demand for more sophisticated chemo-informatics strategies that can streamline data processing and simultaneously maximize disease-relevant molecular information capture. In broad terms, these strategies involve (*i*) raw analytical signal preprocessing for improved information recovery; (*ii*) imaging informatics for correlation of MSI and histological information; and (*iii*) pattern recognition analysis for topographically localized biochemical feature extraction. These steps will influence one another and thus need to be considered within an integrated bioinformatics solution (16).

Typically, data preprocessing methods involve peak detection or "binning" and filtering of solvent/matrix or noise-related peaks (17–19), followed by a normalization step. At present, the most widely applied approach involves integrating MSI spectra within a predefined "bin" size (typically ~0.01 Da). This reduces mass detection accuracy and introduces biologically irrelevant spectral features, making unambiguous assignment of chemical species more difficult. In the case of normalization, the total ion current (TIC) scaling factor is frequently cited in the literature as an acceptable means of accounting for global intensity changes in a MSI dataset (20–22). However, we have recently demonstrated that the performance of this method can be compromised by single large molecular ion peak intensities (21, 23). An additional problem inherent to MS-based analysis of complex biological mixtures is the fact that molecules present in greater intensities within a given sample will tend to exhibit larger variations when subjected to repeated measurement (23). This disruption to variance constancy across the measurement range, known in statistical terms as heteroscedasticity, represents a significant barrier to the effective application of commonly used multivariate techniques for the downstream statistical interrogation of MSI datasets (23). To date a number of different strategies have been proposed in the literature to stabilize variance across the measurement range (24), and we have recently validated several variance-stabilizing normalization techniques in the context of MS-based profiling (23).

Beyond these preprocessing steps, MSI data need to be effectively "fused" with conventional histopathological information to allow the construction of large-scale molecular databases composed of region-specific molecular biomarkers and facilitate future automated histology initiatives. Precise methods for coregistration of histological and MSI data are an essential prerequisite for these applications and represent a further challenge at present. Of the software packages currently available to the MSI analyst, only the proprietary Bruker package offers image coregistration and region-of-interest molecular ion pattern extraction, with the option to further process extracted spectra in the associated statistical toolbox ClinProTools (25). However, this approach (limited to data collected on Bruker instrumentation) requires the user to manually select features on the pre- and poststaining images to conduct coregistration and can be subject to considerable error. Other less refined platforms have sought to achieve this objective by visual selection of particular regions of interest on hematoxylin and eosin (H&E)-stained optical images, followed by selection of pixels occupying similar (but not precisely aligned) geographical coordinates on corresponding MSI heat maps (22, 26). This permits only very crude colocalization of features from the two imaging modalities and may be deemed sufficient perhaps only in instances where limited variation in cell typology is seen across the tissue section (e.g., cancerous cellular regions and healthy cellular regions only). A number of image informatics methods have been recently developed to segment and to align the objects between images (27, 28). These approaches can involve rigid or nonrigid transformation, depending on object deformation characteristics. The most commonly used methods are based on extensions of the Lucas–Canade algorithm and their relative advantages and limitations have been recently described within a unifying framework (27). However, there is no standardized image coregistration protocol in the context of histology-driven MSI, and the currently used marker-based/fiducial methods may lack the precision required for detailed definition of morphology-to-chemistry interrelationships.

Histology-driven, automated tissue identification further requires efficient and robust extraction of tissue-specific molecular ion patterns (19). The multidimensional nature of MSI datasets calls for effective dimensionality reduction techniques that are capable of extracting tissue-specific multivariate molecular ion patterns. Currently, the most widely used supervised dimensionality reduction technique is partial least-squares discriminant analysis (PLS-DA) (29, 30). It has been shown that PLS-based discriminant components are derived by maximizing between-class variance (Table S1) (30). A more mathematically eloquent mode of discriminant analysis is to maximize the difference between class means while simultaneously minimizing within-class variability. This is the objective of linear discriminant analysis (LDA), which maximizes the ratio of between- vs. within-class variance (31). Unfortunately, LDA cannot be directly applied in circumstances where the number of variables exceeds the number of samples, as is the case with the dataset presented here. Principal component analysis (PCA) has been commonly applied as a preprocessing step before LDA (PCA-LDA) to mitigate this problem (32). However, a problem arises here with respect to the selection of an optimal number of components. Introducing too many components into a model will increase the likelihood of LDA model overfit, whereas retaining too few can result in the loss of discriminatory information (33). In the current study, we have proposed the use of a modified maximum margin criterion (Table S1) to improve supervised feature extraction, while simultaneously avoiding arbitrary selection of the number of principal components before discriminant analysis (34).

Here, we have devised a comprehensive data analysis framework with the aim of addressing the current challenges outlined above in MSI data treatment and exploration. Specifically, innovative bioinformatics solutions proposed in this study are *i*) variance-stabilizing normalization for improved information recovery; *ii*) an automated image coregistration algorithm for intuitive, precise histology-to-chemistry feature correlation; and *iii*) a unique method for efficient extraction of tissue-specific multivariate ion patterns. As a validation step, the outlined workflow has been applied to the investigation of tumor-surrounding lipid signatures in colorectal cancer (Movies S1 and S2). We demonstrate that this platform provides in-depth insights into tumor biochemistry by simultaneously analyzing the spatial distribution of hundreds to thousands of lipid species across different cell types. This offers potential for the development of next-generation cancer biomarkers and also may have a translational impact

beyond the field of clinical histopathology in personalized pharmacotherapy and drug discovery.

Results and Discussion

Noise- and Solvent-Related Peak Removal Strategy. To ensure that spectral profiles obtained are of genuine biological relevance, it is essential to filter out spurious peaks related to measurement noise and/or applied experimental solvent(s). Typically, peaks due to noise are present in a small number of randomly selected pixels within a tissue object. This is evident from the distribution of molecular ion features across pixels as illustrated Fig. S1. This distribution shows that hundreds of thousands of molecular ion species are present in <1% of the total number of tissue object-related pixels, whereas imaging technologies are currently capable of detecting an order of magnitude fewer species of biological origin. Based on this observation we have considered peaks to be of biological significance only when found in at least 1% of pixels within the tissue object. A similar strategy has been previously described for MALDI-MSI data analysis (19, 35), although here the authors applied this principle to all pixels in the MSI image, not just those within the confines of the tissue object. The approach used in the present study refines this method further by applying the outlined strategy only to pixels within the tissue object, which should minimize the rate of false-positive genuine peak discovery. In addition, *m/z* species showing higher intensity outside of the boundaries of the tissue area in the image were also discarded. The combination of these processing steps resulted in a significant reduction of spectral data volume from >180,000 to ~1,000–5,000 *m/z* values, thus enabling more efficient data handling/mining.

Log-Based Variance-Stabilizing Normalization for Improved Information Recovery. In the present study we have applied log-based variance-stabilizing normalization (VSN) in favor of other commonly used approaches. The downstream application of multivariate statistical techniques assumes that the MSI measurement noise of molecular levels is consistent across the whole intensity range (23, 24). In the case of MSI datasets these requirements are not met as error structure is characterized by increasing technical variance as a function of increased signal intensity (23). Consequently, molecular species with higher peak intensities exhibit larger variability when repeatedly measured and thus weak signals can be overwhelmed by the noise of strong signals.

To assess the performance of the log-based VSN strategy we selected DESI imaging data from three different tissue regions (Fig. 2*A*), determined by a pathologist to be almost entirely morphologically uniform (i.e., where substantial biological variation is not expected within a given region). Fig. 2*B* illustrates the SD as a function of the rank of the mean of peak intensity across the range of tissue regions. In the absence of heteroscedastic noise structure, the running median of the SD should verge on horizontal, with minor oscillations only but no obvious trend, per se. In Fig. 2*B* this condition is not met and the variation of peak intensity is seen to increase with the rank of mean intensity (i.e., as intensity increases). After application of log-based VSN a clear improvement is seen in the stability of transformed peak intensities across the measurement intensity range (Fig. 2*B*). This ensures that data structure is consistent with the assumption of downstream statistical modeling techniques. This is exemplified in Fig. 2*C*, where PCA has been applied to identify overall similarities/differences in lipid composition across different tissue regions. Before applying the log-based VSN, the resulting PC scores are impacted heavily by the random variation of high-intensity molecular ion peaks and therefore can poorly represent the overall variation structure of a dataset (23).

Implementation of the Image Alignment Algorithm for Precise Coregistration of Morphological and Biochemical Features. In this study, we have sought to profile topographically localized biochemical signatures in many different tissue types (tumor cells, healthy mucosa, connective tissue, smooth muscle, and

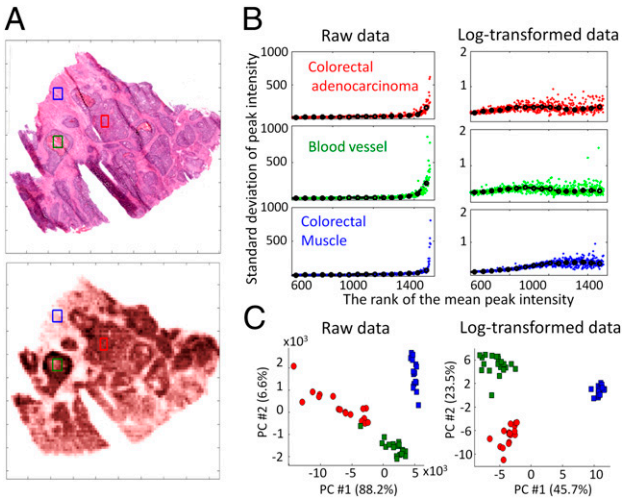


Fig. 2. Impact of variance-stabilizing normalization on information recovery via PCA. (*A*) Coselection of morphologically homogeneous tissue regions. (*B*) SD vs. rank of the mean diagnostic plots for heteroscedastic noise structure. (*C*) PCA before and after variance-stabilizing normalization. Bicross validation has been performed to confirm that PCs capture biologically relevant information not attributable to noise (40).

microvasculature). Moreover, tissue sections were frequently found to harbor multiple different cell types, occupying highly irregular distributions, making more exact feature coregistration not only desirable, but also essential for precise correlation of morphology and biochemistry. For example, Figs. 2 and 3 illustrate how this method enabled highly accurate localized biochemical profiling of various tissue types, including small vascular channels (<5–10 pixels) encased within islands of tumor (Fig. 2*A*).

The image alignment approach used in the present study operates using an in-house–developed automated affine image transformation (translation, rotation, and scaling) algorithmic script based on a gradient descent optimization approach. The coregistration of objects on optical and DESI-MSI images by means of the affine transformation is illustrated in *SI Materials and Methods* (Figs. S2 and S3). The majority of currently available image coregistration algorithms are not fit for the purpose of this setting as they incorporate a manual, fiducial marker-based system that is hampered by user bias and thus lacks the required precision and reproducibility (36). Using our proposed method, once the "template" (optical) and "reference" (DESI-MSI) images have been aligned, the user has the option of using a dual image magnification function for detailed correlation of features from both images in tandem, down to single-pixel resolution. This function proves particularly useful when profiling discrete areas of histological interest (such as isolated clusters of lymphoid tissue, discrete vascular channels, or invading tumor border zones, as exemplified in Figs. 2 and 3).

Enhanced Tissue-Specific Molecular Ion Pattern Extraction via the Recursive Maximum Margin Criterion. Following coselection of multiple regions of interest on optical and corresponding DESI-MSI images (Figs. 2*A* and 3*A*), tissue-type–specific pixel regions (composed of ~20 pixels) are subjected to supervised dimensionality reduction. The objective is to use the least number of components to capture tissue-type–specific chemical signatures based on weighted combinations of molecular ion patterns. Here we have elected to simultaneously maximize the difference between interclass and intraclass variability, using the proposed recursive maximum margin criterion (RMMC) approach. This is a similar objective to that of our previously used PCA-LDA (11). However, the RMMC approach avoids the necessity of selecting an optimal number of PC components before applying LDA, which can result in model underfitting or overfitting. The advantage of

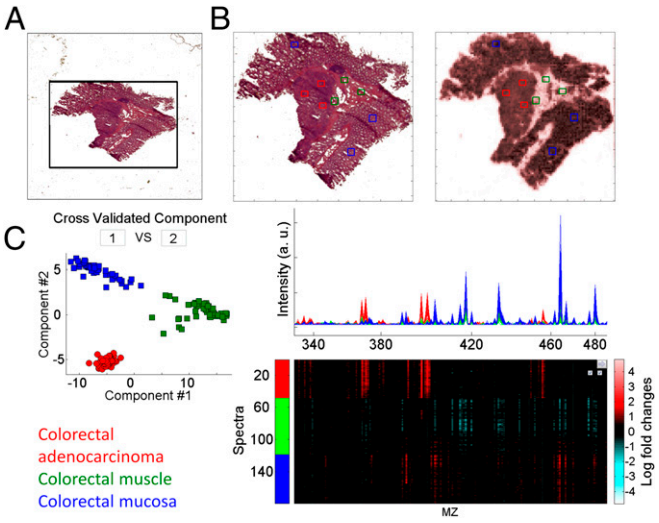


Fig. 3. Image coregistration, feature coselection, and multivariate analysis. (*A*) Automatic image transformation for accurate coregistration of biochemical and histological features. (*B*) High-resolution optical image of H&E tissue section with regions of tumor (red boxes), muscle (green boxes), and healthy mucosa (blue boxes) selected. Shown is aligned DESI-MSI image with automated coselection of pixels corresponding to defined regions of interest. (*C*) Discriminatory analysis using the RMMC method with leave-region-out cross-validation for enhanced separation of tissue classes based on biochemistry.

RMMC over PLS is illustrated using simulated data derived from two multivariate normal distributions with differing means (Fig. S4). The PLS-based discriminating scores result in greater separation of class means with inferior class separability, as shown by the substantial overlap of score distributions between classes. In comparison, the RMMC approach diminishes within-class scatter leading to improved class separability. To further demonstrate the practical advantage of RMMC over PLS-based discriminating feature extraction, we have derived components (the number of classes minus one) for classification of distinct tissue types from two different tissue sections (Table S2). "Leave-region-out" cross-validation with *k*–nearest-neighbor classification has been used. This procedure involves iteratively leaving one randomly designated pixel region out and performing supervised dimensionality reduction on the retained tissue-specific pixels. The left-out pixel regions are then projected onto the trained discriminating space and subsequently subjected to *k*–nearest-neighbor classification. Superior classification accuracy was achieved using RMMC-based derived components, indicating better separability of different tissue types (Table S2).

"One Against All" Tissue-Specific Discriminating Pattern Extraction. The purpose of the final step has been the development of methods that can be used for automated tissue annotation according to biochemistry. It can be envisioned that transition from one cell type to another may not necessarily be associated with abrupt biochemical changes, for example in the case of an infiltrating tumor. This change can be captured by continuous discriminatory scores based on weighted combinations of multivariate ion species from pixels corresponding to a particular tissue type compared with all others. This strategy is known as "one against all" in pattern recognition (37). We have used the RMMC approach to extract tissue-specific molecular ion patterns capable of discriminating histologically defined pixels belonging to one particular tissue type from all others. Subsequently these patterns are used to derive a continuous score for every pixel in the tissue section. These scores can be visualized in a color-coded manner with the color intensity proportional to the score of a given pixel. This is exemplified in Fig. 4, which shows highly accurate chemical reconstruction of different morphological regions [smooth muscle (blue), blood vessels (green), and colorectal cancer (red)] based on continuous scores derived from multivariate molecular ion patterns obtained from representative pixels as described.

The discriminatory scores are then used to generate a "probability" of a given pixel belonging to a given tissue type by means of logistic regression (38) from which the tissue classification is derived, as described in *Materials and Methods*. Pixel-wise tissue-class assignment as a function of the number of training pixels used is illustrated in Figs. S5 and S6. This analysis shows that accurate pixel-wise classification was achieved using just nine reference training spectra per given tissue class. Tissue-type–specific spectra identified in this way can then be exported into a "histologically authentic" MSI database. Tissue-section–specific profiles acquired from cancer-bearing regions can then be subjected to unsupervised multivariate analysis to determine the diversity of cancer-cell metabolic phenotypes for a given case compared with another (41). There are biological circumstances that challenge unambiguous pixel assignment; for example, we have shown that pixels occupying junctional regions between infiltrating tumor border and neighboring tissues have an equally high probability of being classified as either tissue type (Fig. S5). Spectra corresponding to these "border" pixels can be used to investigate the unique biochemical changes that are associated with tumor invasion.

Conclusions

MSI offers a means of chemically mapping the tumor microenvironment intact, avoiding the need for time-consuming and disruptive procedural steps such as laser-capture microdissection. The inherently multidimensional nature of MSI datasets challenges conventional data treatment methodologies and has meant that until now the full potential of this emerging technique has remained unfulfilled. The overarching goal of this study was to create a comprehensive analytical framework (Fig. 5) to enable more effective MSI data analysis for exploration of tumor microenvironmental biochemistry. The presented workflow has been specifically designed to address outlined barriers to effective translational MSI and includes optimized data preprocessing steps, precise image coregistration, and efficient tissue-specific molecular ion feature extraction. The methodology described here permits the visualization of molecular signatures with direct correlation to morphological regions of interest, which can offer

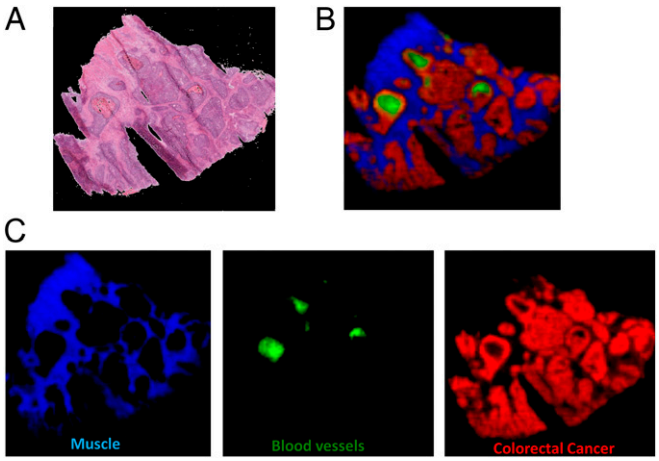


Fig. 4. Chemical reconstruction of tissue regions of interest using multivariate molecular ion patterns. (*A* and *B*) Optical H&E-stained image (*A*) with aligned DESI-MSI RGB image (*B*). (*C*) Reconstruction of three distinct histological regions (smooth muscle, blood vessels, and colorectal adenocarcinoma) based on molecular ion patterns extracted by "one against all" RMMC methodology.

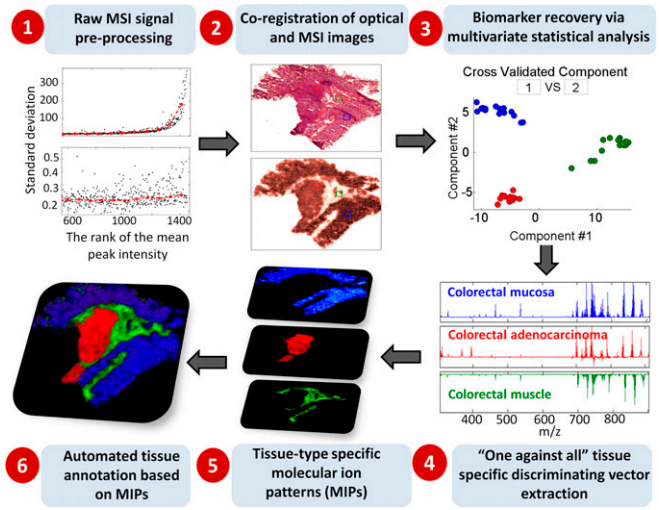


Fig. 5. Overall computational workflow for exploration of region-specific lipid biochemistry using MSI platforms.

unique insights into how different tumor microenvironmental populations interact with one another and generate unique region-of-interest–specific biomarkers and therapeutic targets. In addition, defining morphology based on molecular ion composition allows the development of a histologically authentic MSI spectral database with which to develop highly accurate, fully automated, next-generation tissue classification systems.

Materials and Methods

Tissue Sample Collection and Data Acquisition. This study was approved by the institutional review board at Imperial College National Healthcare Service Trust (Research Ethics Committee reference no. 07/H0712/112). Fresh colorectal cancer tissue samples were harvested from surgical resection specimens in the pathology department and immediately transferred to a freezer at −80 °C before processing. Tissue samples were cryo-sectioned to 10 μm thickness with a SME cryotome (Sigma-Aldrich), using 3% carboxymethyl cellulose as an embedding medium, before being thaw mounted onto plain glass slides. Tissue sections were subjected to negative ion DESI-MSI analysis, using an Exactive mass spectrometer (Thermo Fisher Scientific) coupled with a home-built, automated DESI ion source. The mass resolution used for all measurements was set to 100,000 with a mass accuracy of <4 ppm. Spatial resolution for imaging experiments was set to 75 μm. Methanol/water (95:5 vol/vol) was used as the electrospray solvent at a flow rate of 1.5 μL/min. Zero-grade nitrogen was used as a nebulizing gas at a pressure of 4 bar. Following DESI-MSI analysis, the tissue sections were stained with H&E and digitally scanned at high resolution, using a NanoZoomer 2.0-HT digital slide scanner (Hamamatsu) for precise comparison of biochemical topography and tissue architecture. Further details on data acquisition are provided (*SI Materials and Methods, Tissue Sample Collection and Data Acquisition*).

Data Analysis. The overall workflow starts with optimized preprocessing for improved information recovery. In the case of DESI, this includes classification of tissue object pixels followed by filtering of solvent/noise-related peaks and variance-stabilizing normalization. Next, the tissue object coordinates of MSI and histology images are automatically aligned by means of affine, rigid transformation (translations, rotation, and scaling). The user can then select multiple morphological regions of interest (with zoom functionality) for precise feature coregistration. The spectra from these regions are automatically extracted and subjected to supervised multivariate modeling to derive tissue-specific molecular ion patterns.

Classification of tissue object pixels. The downstream filtering of biologically irrelevant peaks and the application of image coregistration algorithms require accurate separation of tissue object pixels from background. Tissue object pixels usually exhibit higher intensity than background areas, which mainly contain signals due to solvent or noise. Tissue object pixels (*TOP*) can thus be identified by setting an intensity threshold (t) to the total ion intensity (*TI*) image (28); i.e.,

$$TOP(x,y) = \begin{cases} 1 \text{ if } TI(x,y) > t \\ 0 \text{ if } TI(x,y) \le t. \end{cases} \quad [1]$$

Thus, where *TI* at a given (x, y) coordinate exceeds the designated intensity threshold (t), the pixel is classified as *TOP*; otherwise it is assigned to background.

A variety of binary image segmentation algorithms can be used to derive mathematically optimal threshold (t) values (for a comprehensive review of binary segmentation algorithms refer to ref. 28 and references therein). Here, the optimal threshold value was automatically identified by maximizing the difference between tissue object and background, using a histogram-based method as described by Otsu (39). This is arguably the most widely applied global thresholding method for binary image segmentation and compares favorably with other methods (28, 39). To further improve tissue object to background contrast, the *TI* image was calculated disregarding *m/z* features present in the pixel-wide outer border of a section, as these are likely to originate from solvent signals.

Filtering of noise- and solvent-related peaks. Due to inherent inaccuracies in mass detection, molecular ion peaks within an *m/z* range smaller than the native accuracy of the mass spectrometer (<4 ppm in this case) were assigned to the same molecular ion species uniformly for all pixels on a tissue section. The resulting data volume contained hundreds of thousands of noise-related *m/z* species, which are typically located at random in a small percentage of pixels (19, 35). The percentage cutoff value for noise filtration was validated by the systematic analysis of the distribution of molecular ion features across pixels. Based on this validation procedure, molecular ion peaks that were found to be present in <1% of tissue-bearing pixels were removed from subsequent analysis. Additionally, *m/z* species were deemed to be of solvent-related origin if their mean peak intensity within tissue object pixels was less than their mean background intensity.

Variance-stabilizing normalization for improved information recovery. The downstream application of multivariate statistical techniques assumes that measurement noise structure is consistent across the whole intensity range. In the case of MSI datasets this condition is not fulfilled. Here error structure is characterized by increasing technical variance as a function of increased signal intensity, and peak intensities arise through a combination of genuine signals and noise-related signals from different sources, which can be additive or multiplicative in nature (23, 24); i.e.,

$$x_{k(ixj)} = \beta_{k(ixj)} + n_{(ixj)} \cdot s_{k(ixj)} \cdot e^{\eta_{k(ixj)}}, \quad [2]$$

where, $x_{k(ixj)}$ is the measured peak intensity (level) of the kth molecule in the (ixj)th MSI position, $s_{k(ixj)}$ is the expected peak intensity, $\beta_{k(ixj)}$ is the random background noise, $\eta_{k(ixj)}$ is the random multiplicative noise, and $n_{(ixj)}$ is the normalization scaling factor. Additive noise is characterized by random fluctuations in the baseline, irrespective of the presence of molecular signals. Conversely, multiplicative noise grows with the signal intensity of a molecule and is often proportional to it (23). The objectives of variance-stabilizing normalization are to remove biologically unrelated pixel-to-pixel variation in overall signal intensity and to convert multiplicative noise into additive noise for subsequent application of multivariate statistical techniques; i.e.,

$$vsn\left(\frac{x_{k(ixj)}}{n_{(ixj)}}\right) \approx \mu_{k(ixj)} + \varepsilon_{k(ixj)}, \quad [3]$$

where *vsn* denotes a variance-stabilizing normalization, $\mu_{k(ixj)}$ is the transformed peak intensity, and $\varepsilon_{k(ixj)}$ is random additive noise. Here, the normalization factor has been estimated by calculating the median peak intensity, which we have shown is a more robust estimate compared with the widely cited TIC normalization method (21, 23). Due to peak integration and noise-related peak filtration steps, it was assumed that the influence of background noise in the model (Eq. **2**) is negligible. The logarithm was then used as an appropriate variance-stabilizing transformation.

Recovery of tissue-specific molecular ion features via multivariate statistical techniques. The hyperspectral nature of MSI datasets requires effective methods for supervised dimensionality reduction (Table S1). The aim is to derive a series of components composed of weighted sums of molecular ion patterns capable of accurately distinguishing tissue types; *i.e.*,

$$\mathbf{X} = \mathbf{TW} + \mathbf{E}, \quad [4]$$

where **X** is a data matrix after application of variance-stabilizing normalization, **T** is a matrix of derived discriminating components, **W** is a matrix of weights that summarizes the contribution of original variables into discriminating components, and **E** is a residual data matrix. It is assumed that the number of variables is far greater than the number of discriminating

components. Here we propose to maximize the difference (not the ratio) of between- vs. within-class variance (Table S1), which leads to a mathematically stable algorithm, known as the maximum margin criterion (MMC) (34). The original MMC algorithm leads to correlated scores and orthogonal loading vectors. Here, the original algorithm has been modified to derive uncorrelated scores for improved biological interpretation. We have termed this algorithm the RMMC (the pseudocode of this algorithm is summarized in *SI Materials and Methods*).

Coregistration of DESI-MSI and histology images. Image alignment and registration involve denoting the DESI-MSI image as the "reference" and identifying a mathematical (affine) similarity transformation (27) such that the optical H&E-stained ("template") image matches the reference DESI-MSI image as closely as possible with respect to shape, dimensions, and orientation. This type of transformation involves four parameters $[s, \alpha, t_x, t_y]^T$ that denote scaling, rotation, and translations, respectively; i.e.,

$$T = \begin{bmatrix} s\cos\alpha & s\sin\alpha & t_x \\ -s\sin\alpha & s\cos\alpha & t_y \\ 0 & 0 & 1 \end{bmatrix}. \quad [5]$$

The image alignment method proposed here consists of the following steps: (*i*) The optical image pixel resolution is initially decreased to match the pixel resolution of the corresponding MSI image; (*ii*) tissue-related pixels are identified on both optical and MSI images, using the outlined tissue object recognition algorithm; and (*iii*) optimal transformation parameters (scaling, rotation, and translations) for the optical image are found by maximizing the matching of tissue-related pixels between optical and MSI images. The derived transformation is finally applied in full resolution to generate the warped optical image for precise feature coregistration. A variety of optimization techniques can be used to identify the optimal parameter set for affine similarity transformation (for a comprehensive review of image alignment algorithms refer to ref. 27). Here, we have applied the inverse additive algorithm for image alignment (27), which is well suited to the requirements of the present study.

Integrated platform for intuitive processing and visualization workflow. The prototype version of an in-house–developed image coregistration platform has been implemented to allow histology-driven exploration of morphology-specific biochemical features after image alignment, as described above. Using this unique interface, it is possible to zoom in on and select multiple regions of interest and perform multivariate pattern recognition analysis on corresponding spectral profiles (Movies S1 and S2). A built-in discriminatory analytical tool using a one against all approach has been installed into this platform to allow robust separation between different tissue types according to biochemistry. Using this approach, we derive tissue-specific discriminatory scores and multivariate molecular ion patterns by taking one tissue class at a time and comparing it to all others in a given tissue section (hence, one against all). The discriminatory scores are then used to generate a probability of a given pixel belonging to a particular tissue type by means of logistic regression (38). The pixel is assigned to a given tissue class when the derived probability is greater than that expected by random chance ($P > 0.5$). If it is unambiguously assigned to a single tissue class, then these spectra are designated as tissue specific. Fig. S7 demonstrates that morphologically specific classification by these means is robust irrespective of the varying value for the *P*-value designation.

ACKNOWLEDGMENTS. This research project was funded by a Imperial College junior research fellowship (to K.A.V.). The National Institute for Health Research Imperial Biomedical Research Centre, the European Research Council under the starting grant scheme (Contract 210356) and the European Commission FP7 Intelligent Surgical Device Project (Contract 3054940) are also acknowledged for financial support.

1. Chaurand P, Schwartz SA, Caprioli RM (2002) Imaging mass spectrometry: A new tool to investigate the spatial organization of peptides and proteins in mammalian tissue sections. *Curr Opin Chem Biol* 6(5):676–681.
2. Chaurand P, Caprioli RM (2002) Direct profiling and imaging of peptides and proteins from mammalian cells and tissue sections by mass spectrometry. *Electrophoresis* 23(18):3125–3135.
3. Caldwell RL, Caprioli RM (2005) Tissue profiling by mass spectrometry: A review of methodology and applications. *Mol Cell Proteomics* 4(4):394–401.
4. Mirnezami R, et al. (2012) Implementation of molecular phenotyping approaches in the personalized surgical patient journey. *Ann Surg* 255(5):881–889.
5. Nicholson JK, et al. (2012) Metabolic phenotyping in clinical and surgical environments. *Nature* 491(7424):384–392.
6. Elsner M, et al. (2012) MALDI imaging mass spectrometry reveals COX7A2, TAGLN2 and S100-A10 as novel prognostic markers in Barrett's adenocarcinoma. *J Proteomics* 75(15):4693–4704.
7. Schwamborn K (2012) Imaging mass spectrometry in biomarker discovery and validation. *J Proteomics* 75(16):4990–4998.
8. Fletcher JS (2009) Cellular imaging with secondary ion mass spectrometry. *Analyst* 134(11):2204–2215.
9. Passarelli MK, Winograd N (2011) Lipid imaging with time-of-flight secondary ion mass spectrometry (ToF-SIMS). *Biochim Biophys Acta* 1811(11):976–990.
10. Takáts Z, Wiseman JM, Gologan B, Cooks RG (2004) Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science* 306(5695): 471–473.
11. Gerbig S, et al. (2012) Analysis of colorectal adenocarcinoma tissue by desorption electrospray ionization mass spectrometric imaging. *Anal Bioanal Chem* 403(8): 2315–2325.
12. Vickerman JC (2011) Molecular imaging and depth profiling by mass spectrometry–SIMS, MALDI or DESI? *Analyst* 136(11):2199–2217.
13. Marko-Varga G, et al. (2011) Drug localization in different lung cancer phenotypes by MALDI mass spectrometry imaging. *J Proteomics* 74(7):982–992.
14. Prideaux B, Stoeckli M (2012) Mass spectrometry imaging for drug distribution studies. *J Proteomics* 75(16):4999–5013.
15. Kwak JT, Hewitt SM, Sinha S, Bhargava R (2011) Multimodal microscopy for automated histologic analysis of prostate cancer. *BMC Cancer* 11:62.
16. Ghosh S, Matsuoka Y, Asai Y, Hsin KY, Kitano H (2011) Software for systems biology: From tools to integrated platforms. *Nat Rev Genet* 12(12):821–832.
17. McDonnell LA, van Remoortere A, van Zeijl RJ, Deelder AM (2008) Mass spectrometry image correlation: Quantifying colocalization. *J Proteome Res* 7(8):3619–3627.
18. Hanselmann M, et al. (2009) Toward digital staining using imaging mass spectrometry and random forests. *J Proteome Res* 8(7):3558–3567.
19. Alexandrov T (2012) MALDI imaging mass spectrometry: Statistical data analysis and current computational challenges. *BMC Bioinformatics* 13(Suppl 16):S11.
20. Deininger SO, et al. (2011) Normalization in MALDI-TOF imaging datasets of proteins: Practical considerations. *Anal Bioanal Chem* 401(1):167–181.
21. Fonville JM, et al. (2012) Robust data processing and normalization strategy for MALDI mass spectrometric imaging. *Anal Chem* 84(3):1310–1319.
22. Jones EA, Deininger SO, Hogendoorn PC, Deelder AM, McDonnell LA (2012) Imaging mass spectrometry statistical analysis. *J Proteomics* 75(16):4962–4989.
23. Veselkov KA, et al. (2011) Optimized preprocessing of ultra-performance liquid chromatography/mass spectrometry urinary metabolic profiles for improved information recovery. *Anal Chem* 83(15):5864–5872.
24. Rocke DM, Durbin B (1995) Two-component model for measurement error in analytical chemistry. *Technometrics* 37(2):176–184.
25. Ketterlinus R, Hsieh SY, Teng SH, Lee H, Pusch W (2005) Fishing for biomarkers: Analyzing mass spectrometry data with the new ClinProTools software. *Biotechniques* 38(S6):37–40.
26. Eberlin LS, et al. (2013) Ambient mass spectrometry for the intraoperative molecular diagnosis of human brain tumors. *Proc Natl Acad Sci USA* 110(5):1611–1616.
27. Baker S, Matthews I (2004) Lucas-Kanade 20 years on: A unifying framework. *Int J Comput Vis* 56(3):221–255.
28. Sezgin M, Sankur B (2004) Survey over image thresholding techniques and quantitative performance evaluation. *J Electron Imaging* 13(1):146–168.
29. Wold S, Ruhe A, Wold H, Dunn WJ (1984) The collinearity problem in linear-regression - the partial least-squares (PLS) approach to generalized inverses. *Siam J Sci Stat Comp* 5(3):735–743.
30. Barker M, Rayens W (2003) Partial least squares for discrimination. *J Chemometr* 17(3):166–173.
31. Poston WL, Marchette DJ (1998) Recursive dimensionality reduction using Fisher's Linear Discriminant. *Pattern Recognit* 31(7):881–888.
32. Belhumeur PN, Hespanha JP, Kriegman DJ (1997) Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans Patt Anal* 19(7):711–720.
33. Luo D, Ding C, Huang H (2011) Linear discriminant analysis: New formulations and overfit analysis. *25th AAAI Conference on Artificial Intelligence*, eds Burgard W, Roth D. (AAAI Press, San Francisco), pp 417–422.
34. Li H, Jiang T, Zhang K (2003) MMC efficient and robust feature extraction by maximum margin criterion. *Proc Adv Neural Inf Process Syst* 17(1):97–104.
35. Trede D, et al. (2012) On the importance of mathematical methods for analysis of MALDI-imaging mass spectrometry data. *J Integr Bioinform* 9(1):189.
36. Chughtai K, et al. (2013) Mass spectrometric imaging of red fluorescent protein in breast tumor xenografts. *J Am Soc Mass Spectrom* 24(5):711–717.
37. Rifkin R, Klautau A (2004) In defense of one-vs-all classification. *J Mach Learn Res* 5: 101–141.
38. Platt J (1999) Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Adv Large Margin Classifiers* 10(3):61–74.
39. Otsu N (1979) A threshold selection method from gray-level histograms. *IEEE Trans Syst Man Cybern* 9(1):62–66.
40. Owen AB, Perry PO (2009) Bi-cross-validation of the Svd and the nonnegative matrix factorization. *Ann Appl Stat* 3(2):564–594.
41. Fonville JM, et al. (2012) Hyperspectral visualization of mass spectrometry imaging data. *Anal Chem* 85(3):1415–1423.