EVALUATING TUMOR MICROENVIRONMENTS IN THREE DIMENSIONAL CELL CULTURE

MODELS OF COLON CANCER

A Dissertation

Submitted to the Graduate School
of the University of Notre Dame
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

By

Eric M. Weaver

____________________________________
Amanda B. Hummon, Director

Graduate Program in Chemistry and Biochemistry
Notre Dame, Indiana

October 2015
EVALUATING TUMOR MICROENVIRONMENTS IN THREE DIMENSIONAL CELL CULTURE MODELS OF COLORECTAL CARCINOMA

Abstract

By

Eric M. Weaver

Colorectal cancer (CRC) represents an increasingly large threat as it is estimated that 8% of cancer-related deaths worldwide are due to advanced forms of CRC [1]. When detected in early stages, CRC has a very good prognosis upon resection of the primary tumor and continued monitoring of the colon for other aberrant growths. CRC’s biggest threat lies in its ability to readily metastasize to the liver, lymph system and lungs when undetected. Identifying interactions between distinct populations of cells in the tumor architecture that promote metastasis and evade therapy represents a key step in combating this disease.

Using matrix-assisted laser desorption and ionization time of flight mass spectrometry (MALDI-TOF MS), imaging mass spectrometry (IMS) offers the ability to preserve important spatial information in the X,Y and Z planes to observe any analyte
detected over a broad mass range in its native spatial distribution within a sample. Although initially applied to thin sections of mammalian tissues, the application of IMS to sections of three dimensional cell culture tumor mimics has led to a dramatic increase in the ability to study and understand the complex interactions between tumor microenvironments.

The experimental flexibility inherent to cell culture models combined with the robustness of IMS allows for the study of protein expression patterns across the structure of the spheroid, siRNA-mediated knockdown of specific proteins, drug penetration and efficacy studies and lipid expression and localization analysis. The application of chemometrics to IMS datasets has greatly enhanced our understanding of these complex experiments. Overall, these studies have evolved the application of IMS to three dimensional cell culture models into a robust and viable experimental technique that will have an enormous impact on the understanding of the complex interaction between tumor microenvironments and how drug treatment and chemotherapeutics affect tumor microenvironments.

Dedication

This work is dedicated to my family, past, present and future. Without their constant support and encouragement, I would not be where I am today.
# CONTENTS

Figures ................................................................................................................................ vii

Tables ................................................................................................................................... xvii

Acknowledgments........................................................................................................... xviii

Chapter 1: Imaging Mass Spectrometry: From Tissue Sections to Cell Cultures ........... 1
  1.1 Abstract ............................................................................................................. 1
  1.2 Introduction ...................................................................................................... 2
  1.3 Traditional Applications .................................................................................... 7
    1.3.1 Tissue Samples ................................................................................... 7
    1.3.2 Neuronal Samples ............................................................................ 11
    1.3.3 Imaging of Pharmaceutical Compounds .......................................... 11
  1.4 Sample Preparation ........................................................................................ 15
    1.4.1 Sectioning of Biological Samples for IMS ......................................... 15
  1.5 Ionization Approaches for Imaging ................................................................. 18
    1.5.1 MALDI ............................................................................................... 18
    1.5.2 SIMS ................................................................................................. 26
    1.5.3 DESI .................................................................................................. 30
  1.6 Instrumentation .............................................................................................. 33
    1.6.1 TOF Mass Analyzers ......................................................................... 34
    1.6.2 Magnetic Sector Mass Analyzers ..................................................... 35
    1.6.3 Orbitrap Mass Analyzer ................................................................... 35
  1.7 Data Analysis ................................................................................................... 36
    1.7.1 Statistical Analysis ............................................................................ 37
    1.7.2 3D Reconstruction ........................................................................... 39
  1.8 Challenges ....................................................................................................... 42
    1.8.1 Data Analysis .................................................................................... 42
    1.8.2 Spatial Resolution ............................................................................ 43
  1.9 A New Application: Cell cultures ..................................................................... 48
    1.9.1 Imaging Cell Cultures via MALDI ...................................................... 50
    1.9.2 SIMS Imaging of Cell Cultures .......................................................... 53
  1.10 Perspectives .................................................................................................. 54
  1.11 Acknowledgments ......................................................................................... 55
  1.12 References .................................................................................................... 55
4.3 Results and Discussion .................................................................................................................. 128
  4.3.1 Workflow Overview ................................................................................................................. 128
  4.3.2 Isolation and Initial Processing of Relevant Pixels ................................................................. 130
  4.3.3 Cluster Analysis ...................................................................................................................... 135
  4.3.4 PCA-LDA ................................................................................................................................. 138
4.4 Experimental ............................................................................................................................... 147
  4.4.1 Cell Culture ............................................................................................................................. 147
  4.4.2 Sectioning and Sample Preparation for MALDI Imaging ...................................................... 147
  4.4.3 MALDI Imaging ...................................................................................................................... 148
  4.4.4 Data Analysis .......................................................................................................................... 149
4.5 Conclusions .................................................................................................................................... 149
4.6 Acknowledgments .......................................................................................................................... 151
4.7 References .................................................................................................................................... 151

Appendix A: Supplementary Information for Chapter 2: Single Cell Metabolic Profiling of Tumor Mimics ........................................................................................................................................ 155
  A.1 Overview of Glycosphingolipid (GSL) Metabolism ................................................................. 155
  A.2 Cell Diameters ............................................................................................................................ 156
  A.3 CE-LIF Analysis of Cellular Incubation Medium ...................................................................... 157
  A.4 Three-Color Single Cell Electropherograms ......................................................................... 158
  A.5 Effect of Chloroquine on Fluorescent GSL Uptake ............................................................... 161
  A.6 Percent Relative Peak Areas for all Fluorescent GSL Species ............................................. 162

Appendix B: Chemometric analysis of MALDI Mass Spectrometric Images of Three Dimensional Cell Culture Systems: MATLAB Processing User Guide .................................................................................................. 166
  B.1 Part One: PCA of Relevant Pixels ............................................................................................... 168
    B.1.1 Load Data (Optional) ........................................................................................................... 168
    B.1.2 Trim Data (Optional) .......................................................................................................... 170
    B.1.3 Gaussian Smooth Data (Optional) ...................................................................................... 171
    B.1.4 PCA on All Pixels ................................................................................................................ 176
    B.1.5 Isolation of Relevant Pixels ................................................................................................. 176
    B.1.6 Pre-Processing of Selected Subset of Data ......................................................................... 178
    B.1.7 Align and Area-Normalize ................................................................................................. 180
    B.1.8 Perform PCA on the Processed Data of Relevant Pixels ..................................................... 183
    B.1.9 Plot the Principal Components and Score Images (Optional) ........................................... 186
  B.2 Part Two: Clustering .................................................................................................................... 188
    B.2.1 Estimation of Cluster Number (Optional) .......................................................................... 188
    B.2.2 Cluster Analysis .................................................................................................................... 190
  B.3 Part Three: PCA-LDA .................................................................................................................. 193
    B.3.1 Performing PCA-LDA and Obtaining a Discriminant Spectrum ........................................ 193
  B.4 References .................................................................................................................................... 197

Appendix C: Chemometric Analysis of MALDI Mass Spectrometric Images of Three-Dimensional Cell Culture Systems .......................................................................................................................... 198
FIGURES

Figure 1.1 An overview the general IMS workflow for 3D cell cultures. Cells are plated in 96-wells with a layer of agarose dissolved in cell culture media at the bottom of the inner 60 wells to facilitate 3D spheroid formation. After 10-14 days the spheroids grow to roughly 1 mm in diameter, which is large enough to establish nutrient and oxygen gradients mimicking tumor formation in vivo. The spheroids are then collected and embedded in gelatin for sectioning and subsequent IMS analysis. The workflow preserves spatially relevant proteomic information and enables examination of the distribution of proteins without the use of labels or pre-selection of analytes. ROI (Region of Interest) 1 corresponds to the spatial distribution of the proteins species in the necrotic core while ROI 2 delineates the outer perimeter of the spheroid sample.

Figure 1.2: Demonstration of detection of the pharmaceutical compound olanzapine (OLZ) in a whole body section of a male rat 2 hours post-dosage. A) Optical image with the organs highlighted in pink. B) Distribution of an MS/MS ion representing the m/z value 256 Da showing the localization of OLZ within the rat. C) Localization of the MS/MS ion species m/z 256 that displays the distribution of the N-desmethyl metabolite of OLZ. D) MS/MS ion image of m/z 272, the 2-hydroxy methyl metabolite. Species were identified using the MS/MS capabilities of the TOF mass analyzer that breaks molecules into small component pieces and identifies them based on their fragmentation patterns. The species were identified during the experiment and not preselected or labeled prior to analysis. Scale bar = 1 cm Reproduced with permission from reference [22].

Figure 1.3: Imaging Mass Spectrometry analysis of a mouse epididymis, a highly complex, compartmentalized organ with distinct areas containing different protein species that can be tracked via their distinct m/z values. A) H&E stain of a tissue section after analysis with discrete regions annotated to the right of the picture. B) Averaged mass spectrum from all of the pixels from the tissue image. C) False color ion intensity maps with each color corresponding to m/z values derived from different proteins in the compartments of the tissue. Reproduced with permission from reference [32].

Figure 1.4: Example of a multi-modal imaging experiment combining traditional H&E staining, DESI imaging mass spectrometry and MALDI imaging mass spectrometry. A through C are the ion intensity maps from different lipid species.
imaged using a DESI source, D through F are protein species imaged via MALDI mass spectrometry and G through I are H&E stains. Each method was carried out on the same tissue slice of a mouse brain. The combination of information from each method provides a unique, complementary perspective by allowing analysis with each technique. Reproduced with permission from reference [50].

Figure 1.5: High resolution MALDI IMS of rat testes performed with a 20 µm laser raster. A) H&E stain of two compartments of the tissue denoted by the Roman numerals in the image. B) MALDI false color ion image for four different protein species precisely detected in each compartment. This experiment demonstrates a large step towards improved lateral spatial resolution in MALDI IMS experiments. Scale bar is 200 µm. Adapted with permission from [99].

Figure 1.6: Depiction of 4 protein species detected on a slice from a tumor spheroid model grown from the colorectal carcinoma cell line HCT 116. Each species is detected in a spatially defined region of the spheroid and m/z value 12,828 is clearly localized to the inner portion of the spheroid slice. Detection of this species highlights the biologically relevant gradients recapitulated in 3D cell culture models by demonstrating proteomic differences between the inner necrotic core and other regions of the tumor spheroid mimicking tumor formation in vivo. Reproduced with permission from reference [1].

Figure 1.7: An example of high resolution SIMS IMS on multiple cell culture substrates. A through C are phase contrast images of the cells in culture prior to SIMS imaging and D through F are the corresponding total ion maps from SIMS analysis. A and D are rat esophageal cells, B and E are NIH 3T3 fibroblast cells and C and F are a heterogeneous culture consisting of both rat esophageal and 3T3 cells. This figure demonstrates both the application of coculturing to provide a more accurate cell culture model of the in vivo environment and the capability of SIMS IMS to provide extraordinarily high-resolution images. It also demonstrates the ability of SIMS to identify different cell types in the same cell culture substrate. Scale bar is 100 µm. Reproduced with permission from reference [69].

Figure 2.1: Experimental overview. HCT 116 MCTSs were simultaneously incubated with GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665. After 24h, the spheroids underwent serial trypsinization, yielding sheet of cells from the outer (orange), middle (magenta), and core (cyan) regions. Dead cells and cellular debris are contained within the center (gray); this material was discarded. These cellular sheets underwent trituration to isolate individual cells, and the cells were then fixed with paraformaldehyde (PFA). After fixation, cells were analyzed with CE-LIF.

Figure 2.2: Electropherograms from single cells of HCT 116 MCTSs. A, B, and C show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the outer region of HCT 116 MCTSs.
Numbers indicate the metabolic product (1-GM1, 2-GM2, 3-GM3, 4-LacCer, 5-GlcCer, and 6-Cer). Red numbers indicate the fluorescent substrate for each channel.

Figure 2.3: Average multicolor metabolic cytometry electropherograms, normalized to unit area, for each region within the spheroids. The line width encompasses + standard error of the mean for all cells in each location and pharmacological condition. A, B, and C compare GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) by location in HCT 116 MCTS single cells. D, E, and F present enlargements of the Y-axis of A, B, and C (respectively). G, H, and I compare by location in chloroquine (CQ)-treated MCTS. J, K, and L present enlargements of the Y-axis of G, H, and I (respectively). Numbers indicate the metabolic product (1-GM1, 2-GM2, 3-GM3, 4-LacCer, 5-GlcCer, and 6-Cer). Red numbers indicate the fluorescent substrate in each channel. An asterisk indicates a significant difference (p < 0.05) in peak area using a One-Way ANOVA as described in the text.

Figure 2.4: Effect of chloroquine treatment on fluorescent GSL metabolism by region. The data in Figure 3 are recompiled to illustrate the effect of chloroquine (CQ) in different regions. Each colored trace is shaded to encompass + standard error of the mean for all cells in a specified location and pharmacological condition. A, B, and C show the effect of CQ on GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells from HCT 116 MCTSs located within the outer region. D, E, and F show the effect of CQ on GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) within the middle region. G, H, and I show the effect of CQ on GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) within the core region. Numerical indices are the same as those depicted in Figure 3. An asterisk indicates a significant difference (p < 0.05) in peak area using a Student’s t-test as described in the text.

Figure 2.5: Principal component analysis of multicolor single cell electropherograms. A. PC1 versus PC2 score plot of the concatenated multicolor electropherograms for each cell in each condition. B. Plot of the Euclidean Norm of PCA scores between each point within a specified cluster and the multivariate spatial median for all points within the specified cluster. Error bars are ± one standard error of the mean.

Figure 3.1: Conceptual overview of “Spatial Metabolomics Pipeline” as applied to complex high resolution IMS datasets. Adapted from Dr. Andrew Palmer.

Figure 3.2: Metabolic trends from different spheroids at different time points of chloroquine treatment. 706.5381 m/z corresponds to a completely unsaturated lipid species. The data indicates that this particular lipid species is most likely not being used an energy source in the outer population of cells at each time point.
Conversely, 766.5745 \textit{m/z} indicates a partially saturated lipid species. The localization of this species implies that the inner populations of cells in the spheroid at each time point are being forced to rely on \( \beta \)-oxidation as their primary energy source due to lack of other metabolites. There is a considerable increase in the amount of saturated lipid species in the core of 24 hour spheroid.104

Figure 3.3: Carnitine Distribution in two different untreated spheroids at similar longitudinal depth in each spheroid. Carnitine is shown to be distinctly located in the intermediate region of these spheroids. This data could signify an upregulation of \( \beta \)-oxidation pathways in these regions. The multiple types of carnitine present likely indicate different lipid species being involved in these pathways. ................................................................. 107

Figure 3.4: Chloroquine distribution (found at 320.1887 \textit{m/z}, left column) in an untreated, twelve hour treated and 24 hour treated spheroid sectioned at roughly the same longitudinal depth in each spheroid. Imaged in the right column is the diprotonated form of chloroquine (found at 322.2043 \textit{m/z}). ..... 108

Figure 3.5: Workflow overview of “Spatial Metabolomic Pipeline.” The pipeline combines IMS data, database searches and conventional MS/MS sequencing data to assist in the analysis of highly complex datasets. Adapted from work done by Dr. Andrew Palmer............................................................... 110

Figure 3.6: The top row of boxes shows the JavaScript rendering of the distribution of HMDB08915. Represented are the Sodium adduct of PC(18:3(6Z,9Z,12Z)/P-18:1(9Z)) that is localized heavily to the core of 0 hr (Untreated) and 12h treated spheroids. The bottom row of boxes shows the distribution of the molecular ion corresponding to HMDB07876 (PC(14:0/18:3(9Z,12Z,15Z))) which shows a variable distribution across time points. .............................................................................. 111

Figure 3.7: Log fold change compared to the Untreated samples of HMDB08915 demonstrating the magnitude of abundance differences between the treated time points. .................................................................................. 112

Figure 3.8: Summed MS level spectra from one replicate of the direct injection ESI-MS/MS analysis of each population of cells from untreated spheroids. Insets are MS/MS spectra and the structures of the lipids they represent as well as the structural analysis obtained by submitting the search results to the LIPID MAPS online database.......................................................................................... 114

Figure 3.9: Schematic visualization of serial trypsinization procedure demonstrating the separation of the biologically relevant regions of a fully grown HCT 116 CRC spheroid. ............................................................................................................. 115
Figure 3.10: Visualization of direct injection ESI data shown in Figure 3.9 depicting the spatial distribution as well as the relative intensity in each region of three lipid species based on normalized intensities of each lipid parent ion detected. Peak intensities were normalized and averaged across three replicates collected from untreated spheroids. Adapted from work done by Dr. Andrew Palmer. ........... 116

Figure 4.1: Summary of data workflow. MALDI IMS data is first imported into the MATLAB programming environment. PCA is then used to isolate relevant pixels within the image. The mass spectra of the isolated pixels are then further processed before a second round of PCA that highlights general spectral features associated within the sample. Clustering analysis is then performed using the scores of the relevant PCs from this second iteration of PCA to reveal the spatial distribution pattern of distinct cellular populations. Finally, the results of both the PCA and clustering of relevant pixels are used as inputs into an LDA routine that identifies specific \( m/z \) values selectively associated within each of the distinct cellular microenvironments in the spheroid. ........................................ 129

Figure 4.2: Isolation of Relevant Pixels and Spectral Processing. A) Example mass spectrum of a pixel discarded after an initial round of PCA on the unprocessed data. B) Isolation of mass spectra from pixels associated with spheroid cells (white) using PCA. Spectra arising from pixels outside of the sample (grey) were discarded. C) The total ion spectrum of all relevant pixels identified in B (white pixels). D) The total ion spectrum after log transformation and baseline removal. 132

Figure 4.3 PCA of Isolated Spectra. Principal component loadings (left) and spatially-localized pixel score plots (right) for principal components 1 (top), 3 (middle), and 6 (bottom). ................................................................................................... 135

Figure 4.4 Cluster Analysis. A) Regions of interest identified by HCA. B) The average processed spectra of each region in A. The purple and green traces correspond to the average processed spectra of all purple and green pixels, respectively. C) Regions of interest identified by \( k \)-means clustering. D) The average processed spectra of each region in C. The purple, green, and blue traces correspond to the average processed spectra of all purple, green, and blue pixels, respectively. . 138

Figure 4.5: PCA-LDA Analysis. A and B show score plots of the clustered pixels and the calculated linear boundaries. A) The calculated linear boundary separates the pixels associated with the outer and middle regions of the spheroid. B) The calculated linear boundary separates the pixels associated with the middle and core regions of the spheroid. C) The calculated discriminant spectrum differentiating between the proliferative outer region (shaded blue) and the quiescent middle region (shaded pink). D) The calculated discriminant spectrum differentiating between the quiescent middle region (shaded blue) and the necrotic core region (shaded pink). The lower case Roman numerals highlight several \( m/z \) values discussed in the text. E) Single ion heat maps of \( m/z \) values
identified using the discriminant spectra. The m/z value and its corresponding Roman numeral notation is inset within each two-dimensional heat map. .... 142

Figure A.1: Simplified overview of GSL metabolism in colon cancers......................... 155

Figure A.2: Phase-contrast micrograph of isolated cells. This image was taken of single cells isolated from the outer region of HCT 116 MCTSs. The scale bar represents 25 µm. ................................................................................................................. 156

Figure A.3: Three-Color Electropherogram of the Fluorescent GSL Substrates. The three-color electropherogram represents a 5 nM dilution of the fluorescent GSL substrate-containing incubation medium. ......................................................... 157

Figure A.4: Electropherograms from Isolated Single Cells of HCT 116 MCTSs A, B, and C show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the outer region of HCT 116 MCTSs. D, E, and F show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the middle region of HCT 116 MCTSs. G, H, and I show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the core-adjacent region of HCT 116 MCTSs. Numerical indices indicate the metabolic product (1-GM1, 2-GM2, 3-GM3, 4-LacCer, 5-GlcCer, and 6-Cer). Red numbers indicate the fluorescent substrate in each channel. ............................................................................................................... 159

Figure A.5: Electropherograms from Isolated Single Cells of Chloroquine-Treated HCT 116 MCTSs. A, B, and C show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the outer region of HCT 116 MCTSs treated with chloroquine (CQ). D, E, and F show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the middle region of HCT 116 MCTSs. G, H, and I show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the core-adjacent region of HCT 116 MCTSs treated with CQ. Numerical indices indicate the metabolic product (1-GM1, 2-GM2, 3-GM3, 4-LacCer, 5-GlcCer, and 6-Cer). Red numbers indicate the fluorescent substrate in each channel............................................................................................................... 160

Figure A.6: Total Fluorescence Peak Areas as a Measure of Fluorescent GSL Uptake in isolated HCT 116 MCTS Cells. A, B, and C show a plot of the total fluorescence signal measured within the BODIPY-FL, BODIPY-TMR, and BODIPY-650/665 channels (respectively) for the 167 single cell electropherograms excited at decreased laser powers. The total fluorescence signal within each trace was calculated by summing the peak areas for each metabolite. Asterisks indicate whether chloroquine (CQ) significantly increased the total fluorescence signal.
One, two, and three asterisks indicate p < 0.05, p < 0.01, and p < 0.001, respectively, calculated using a Student’s t-test. The n.s. notation indicates no significant difference. Averages are shown with error bars representing standard error of the mean.

Figure A.7: Average Percent Relative Peak Area for Each Metabolite within the BODIPY-FL Channel as a Function of Location and Pharmacological Treatment. A, B, C, D, E, and F illustrate average percent relative peak areas for GM1-, GM2-, GM3-, LacCer-, GlcCer-, and Cer-BODIPY-FL (respectively). Structures for each metabolite are given above each graph with the ceramide region highlighted blue to indicate the presence of the BODIPY-FL tag. GM3 is underlined to indicate that it was the fluorescent substrate. Averages are shown with error bars representing standard error of the mean.

Figure A.8: Average Percent Relative Peak Area for Each Metabolite within the BODIPY-TMR Channel as a Function of Location and Pharmacological Treatment. A, B, C, D, E, and F illustrate average percent relative peak areas for GM1-, GM2-, GM3-, LacCer-, GlcCer-, and Cer-BODIPY-TMR (respectively). Structures for each metabolite are given above each graph with the ceramide region highlighted green to indicate the presence of the BODIPY-TMR tag. GM1 is underlined to indicate that it was the fluorescent substrate. Averages are shown with error bars representing standard error of the mean.

Figure A.9: Average Percent Relative Peak Area for Each Metabolite within the BODIPY-650/665 Channel as a Function of Location and Pharmacological Treatment. A, B, C, D, E, and F illustrate average percent relative peak areas for GM1-, GM2-, GM3-, LacCer-, GlcCer-, and Cer-BODIPY-650/665 (respectively). Structures for each metabolite are given above each graph with the ceramide region highlighted red to indicate the presence of the BODIPY-650/665 tag. LacCer is underlined to indicate that it was the fluorescent substrate. Averages are shown with error bars representing standard error of the mean.

Figure B.1: Example comparing two variations on using Gaussian smoothing to remove noise from mass spectrometry data. Left graph shows an entire summed spectrum while the graph on the right is a close up of the region between 9000 m/z and 9900 m/z units. In both examples the blue trace indicates the original, non-smoothed data while the red trace represents the summed spectra after the smoothing algorithm has been applied.

Figure B.2: Comparison of summed spectra before and after downsampling is applied.

Figure B.3: The left-hand frame shows the isolation command which allows the user to select which pixels to keep for the next steps of the processing (surrounded by yellow circle). The right hand side shows a representative image of the spheroid being processed in the MATLAB environment. Area shaded in white corresponds
to the pixels highlighted in yellow in the selection step (pictured left). Gray pixels are discarded from the data set and not processed any further in the workflow.

Figure B.4: An example graph produced by this command is shown below that is used to assess the top hat filter only. If the legend is in the way, use the white mouse button to move it elsewhere. Note that the two traces that are viewed are the log-transformed original data prior to top hat filtering (blue) and the log-transformed data after processing the top hat filter.

Figure B.5: Top graph shows all of the spectra plotted on top of one another over the entire mass range to be analyzed. The bottom panels demonstrate zooming in on specific peaks that occur in all of the spectra. When the “align_and_normalize” command is performed all of the spectra will be aligned to the center of the two peaks shown in the bottom panel.

Figure B.6: Graph displayed following proper execution of the “align_and_normalize” command.

Figure B.7: Example of a Scree plot zoomed into the first region. Arrows indicate “elbow regions” representing changes in the variance described by each principal component. This method can be used to estimate the number of principal components that contain actual experimental variation and will be retained in the subsequent processing steps.

Figure B.8: The spectral trend is shown on the left and the heat map of the spectral trend is on the right. The color scale represents the magnitude of the scores of this particular principal component for each pixel; large positive values indicate a positive correlation with a particular pixel and large negative values indicate a negative correlation. Note: In the heat map, the deepest darkest blue areas indicate pixels that were not isolated by the plotandselect command.

Figure B.9: Example of inverting the principal component. The heat map also inverts. This is an example of the “cleangraph” output feature of this command.

Figure B.10: Graph will be generated that displays dispersion as a function of cluster number for the reference and test data sets as shown above. See [4] for more information.

Figure B.11: The graph on the left displays the average processed spectra for pixels of a specific group. The legend is numbered data1, data2, data3... to identify the cluster number associated with each average trace. The graph on the right displays an XY coordinate map of the spatial location of each type of pixel. Note: The color bar identifies which region corresponds to which cluster number. In this example, blue identifies the pixels associated with cluster number 1, green identifies the pixels associated with cluster number 2, red identifies the pixels...
associated with cluster number 3, and black represents irrelevant pixels not
lassoed from the plotandselect command and those not originally imaged (0th
cluster). ............................................................................................................... 192

Figure B.12: The graph on the left shows a cluster plot of the principal component
scores for each group identified. The dark line is the linear
discriminant/boundary calculated by PCALDA. Note that its position successfully
segregates the two clusters of data. The graph on the right shows the
discriminant spectrum. When ‘colored’ mode is used, positive values will be
enclosed within an area of blue while negative values will be enclosed in red. The
line through the center marks zero (this will only work for mean-centered data).195

Figure B.13: Graphical outputs of the “discrimspectra” command. As expected, the
linear discriminant shifts to separate groups 2 and 1. Note that the discriminant
spectrum is not shaded because the ‘singletrace’ distinction was used in the
command line. The dotted line is used to mark the zero position. In either type of
graph, the user may wish to use the zoom tool to identify peaks within the
calculated discriminant spectra. ................................................................. 196

Figure C.1: Example of PCA. Consider a spheroid slice analyzed by MALDI IMS. A mass
spectrum taken from one pixel of a MALDI IMS data set (black, left) can be
represented by a set of spectral features (colored, right) after performing PCA.
PCA identifies several spectral features termed principal components (PCs)
associated within all pixels of this particular data set (red, blue, green, and
magenta graphs). A set of scores (numbers inset within each graph) describe
how important each of these principal components are to the mass spectrum of
this particular pixel. The original data can be reconstituted by multiplying each
principal component loading by its score and summing the result. .................. 200

Figure C.2: Noise Removal by PCA. The original spectrum (black trace) can be
reconstructed nearly identically using only the first several principal components
that describe relevant information (red trace). The latter principal components
describe only noise (purple trace) and can be discarded.............................. 202

Figure C.3: k-means Clustering. A set of data was generated that contained two clusters
of points with random noise added (A). The algorithm randomly chooses two
cluster seeds to start the process (circled red and blue in A). All other points in
the data set are assigned to the nearest cluster (B). Using these associations,
new cluster centers are calculated, represented by the two Xs in B. The process
of assigning points to clusters is repeated again as shown in C. Note how the
cluster assignments changed. The calculation of cluster centers and point
assignments is repeated until convergence is reached (D). ............................ 206

Figure C.4: Calculation of a Linear Boundary by LDA. The data shown here is taken from
Figure C.2.1(D). LDA calculates the equation of a boundary (black line) that
segregates the coordinate system into two regions (orange for cluster 1 and green for cluster 2). Any future observations would be categorized into a specific group depending on its location on either side of the line. ........................................ 209
TABLES

Table 1.1 Commonly used MALDI matrices, their molecular masses, absorbance wavelengths and application ................................................................. 23

Table 1.2 Comparison of IMS approaches detailing ionization methods, mass analyzers, optimal analytes, mass range and resolution ................................................... 26
I would first like to acknowledge my advisor, Dr. Amanda B. Hummon, for being constantly encouraging and supportive throughout my time at Notre Dame. Thank you for being approachable and encouraging of my ideas, patient with my obsession with minutia and for nurturing my growth as a scientist and as an “analytical chemist.”

Thank you to my lab mates, Dr. Kerry Bauer and Dr. Susan “Carl” Skube. Kerry, you always provided a quiet, calm, excellent example as a scientist, and Carl, you were there to help with not only my science questions but also with ‘90s Adam Sandler movie and SNL quotes. Thank you for both.

Of course, I would never have gotten the opportunity to be here without the constant support and opportunities provided by parents. My mom who nurtured my education and love of all things science and nature from a very early age and with liberal doses of NPR and many, many trips to the St. Louis Science Center, and my dad who always set a great example in all matters inside and outside the classroom and through many character building seasons as the coach of my various sports teams. Thank you Bunches for always being there for me with unfailing love and support.

I would also like to thank my extended hockey family (and Andy) who helped distract me from lab work at every chance they could as well as my high school biology
teacher Mr. Talken. Through knowing nothing about biology but a lot of things about teaching 15 year olds, he managed to get us to actually learn something about science. Finally, an enormous thank you to Mass Spectrometry and Proteomics Facility, without their help, advice and patience with all my questions I would definitely have quit.
CHAPTER 1:

IMAGING MASS SPECTROMETRY: FROM TISSUE SECTIONS TO CELL CULTURES

This chapter presents a comprehensive review of the applications of imaging mass spectrometry to cell culture models in general, with an emphasis on moving towards adapting these methods to three dimensional cell culture systems. This chapter was originally published in *Advanced Drug Delivery Reviews* as part of a special theme issue on “Delivery of biopharmaceuticals: Advanced analytical and biophysical methods.”

1.1 Abstract

Imaging Mass Spectrometry (IMS) has been a useful tool for investigating protein, peptide, drug and metabolite distributions in human and animal tissue samples for almost 15 years. The major advantages of this method include a broad mass range, the ability to detect multiple analytes in a single experiment without the use of labels and the preservation of biologically relevant spatial information. Currently the majority of IMS experiments are based on imaging animal tissue sections or small tumor biopsies. An alternative method currently being developed is the application of IMS to three-dimensional cell and tissue culture systems. With new advances in tissue culture and engineering, these model systems are able to provide increasingly accurate, high-
throughput and cost-effective models that recapitulate important characteristics of cell and tissue growth in vivo. This review will describe the most recent advances in IMS technology and the bright future of applying IMS to the field of three-dimensional cell and tissue culture.

1.2 Introduction

Imaging Mass Spectrometry (IMS) is a rapidly developing technology that offers a unique perspective on biological systems. IMS can be applied to virtually any animal or human tissue and, more recently, to model systems grown in vitro [1]. The basic workflow of IMS is depicted in Figure 1.1 and requires that the tissue be thinly sliced (10-20 μm), washed to remove debris and, depending on the ionization method, coated in a thin layer of matrix that assists in the ionization of proteins, peptides, lipids, drugs, xenobiotics, neuropeptides and drug metabolites [2-6]. While there are several ionization sources that can be used, the principle is the same for each: the ionization source raster across the sample in an X,Y coordinate fashion while a mass spectrometer records a mass spectrum for each position. Each “pixel” in the resulting image is a mass spectrum from that specific coordinate position on the tissue. Data analysis programs combine this information with an optical image of the tissue and generate heat maps showing the spatial localization of any m/z value that was measured by the mass spectrometer. This technique allows for visualization of theoretically any ionizable species in a selected mass range while retaining spatial information. A single pixel in an IMS image contains all of the ionized molecules detected for a selected mass range.
Depending on several variables, this could be tens or hundreds of ions/pixel with hundreds or thousands of pixels/image. A wealth of information is obtained from IMS experiments in the form of thousands of analyte signals being detected for each image. Image processing techniques that help highlight biologically relevant signals will be discussed in later sections of this review. IMS images can also be combined with other imaging techniques to provide additional, complimentary information about the samples. IMS can be thought of as “molecular histology” because acquired images can be correlated with H&E stains or other target-based immunostaining methods, and by themselves provide information within a spatially defined context [7].
Figure 1.1 An overview the general IMS workflow for 3D cell cultures. Cells are plated in 96-wells with a layer of agarose dissolved in cell culture media at the bottom of the inner 60 wells to facilitate 3D spheroid formation. After 10-14 days the spheroids grow to roughly 1 mm in diameter, which is large enough to establish nutrient and oxygen gradients mimicking tumor formation *in vivo*. The spheroids are then collected and embedded in gelatin for sectioning and subsequent IMS analysis. The workflow preserves spatially relevant proteomic information and enables examination of the distribution of proteins without the use of labels or pre-selection of analytes. ROI (Region of Interest) 1 corresponds to the spatial distribution of the proteins species in the necrotic core while ROI 2 delineates the outer perimeter of the spheroid sample.
One of the major advantages of the approach is the ability to discern the distribution of previously unknown molecular species. Unlike immunological techniques, IMS is a label-free technique and like all flavors of mass spectrometry (MS), the sequencing of analytes through MS/MS fragmentation can be used to identify unknown biological macromolecules [8]. Importantly, mass spectrometers acquire information about the samples in a chemically specific manner, i.e. the molecular mass and elemental or isotopic state of the analyte directly. The information obtained from an IMS experiment is derived from the analytes and not from an indirect measurement such as a secondary antibody or fluorescent tag [9]. Additionally, there are several labeling techniques that are routine to MS that use isobaric tags or heavy and light isotopes to allow direct comparison and even quantification of proteins from different sample [10, 11].

IMS experiments can be either discovery-based or targeted in design. In “discovery mode,” MS/MS sequencing of analytes detected during the experiment occurs by measuring the mass of a whole molecule and subsequently fragmenting and detecting the component pieces of that molecule. By piecing together the masses of each fragment, researchers can deduce structural information from the analyte such as amino acid composition of proteins, acyl chain length for fatty acids, phosphorylation state of signaling molecules and specific functional groups of organic molecules, to highlight just a few examples [10, 12]. In this way a confident identification of the previously unknown analyte can be made. In “target mode,” the known molecular masses of specific molecules can easily be found and their distributions mapped using
the aforementioned ion heat maps. Generally, due to sample preparation requirements for different classes of molecules, each experiment is limited to detecting only one class of compounds but different types of molecules can be analyzed on adjacent samples and the results combined to give a detailed picture of the interactions between the different kinds of molecules present. While operating in “discovery mode,” IMS requires no prior knowledge of specific analytes to be detected other than the desired class of molecules for a particular experiment. The major advantage of any discovery-based IMS experiment is highlighting targets that were not previously known and combining the information from the IMS experiment and complementary techniques to further discern the relevance of that particular analyte [13].

Due to the increasing popularity of IMS, there have been several quality reviews on the topic in the last few years, providing an overview of the field [14], describing statistical concerns [15], the mapping of phospholipids [16], providing discussions of sample preparation strategies [17], or describing a specific application of the approach [18-20].

The first successful demonstration of matrix assisted laser desorption and ionization (MALDI) imaging was performed by Caprioli et al. in 1997, in which a rat pituitary gland and rat pancreas were imaged using MALDI MS [21]. For MALDI IMS, the analysis of proteins is generally performed in positive ion mode. Samples are prepared with an acidic matrix and proteins are detected with an additional proton, as [M+H]+ species where M corresponds to the molecular mass of the analyte species in Daltons (Da) and the additional proton adds a +1 charge.
In the Caprioli study, one analyte was detected selectively at the edge of an islet in a section of the rat pancreas. The mass to charge ratio (m/z) for this analyte is 5802, which corresponds to the [M+H]+ value for insulin. The differences between the islet and other parts of the tissue were examined. Detection of insulin at m/z 5802 was compared to a negative control signal at m/z 6500. The results showed the distribution of m/z 5802 matched previous size estimates for a rat pancreas islet, whereas the S/N for m/z 6500 showed no change above background intensity in the region of interest, allowing accurate mapping of the islet based on the m/z values observed.

Since these first experiments, IMS analyses have advanced in all aspects, from sample preparation to instrumentation and data analysis. This review will cover the basics of IMS and end with a discussion of a novel, growing application of the technique; imaging cell culture systems.

1.3 Traditional Applications

1.3.1 Tissue Samples

Technological advances of all aspects of the IMS workflow enables imaging virtually any animal tissue, whole body sections, a variety of 2D cell culture models and 3D cell culture models [1, 4, 5, 22, 23]. The detection of both drug and metabolite species in whole body rat sections is shown in Figure 1.2[22]. Detailed studies of a wide variety of disease states have been done using IMS including analysis of Parkinson’s and Alzheimer’s diseases, breast, colon, prostate, lung and pancreatic cancer and various forms of gliomas [1, 5-7, 24-28].
Figure 1.2: Demonstration of detection of the pharmaceutical compound olanzapine (OLZ) in a whole body section of a male rat 2 hours post-dosage. A) Optical image with the organs highlighted in pink. B) Distribution of an MS/MS ion representing the m/z value 256 Da showing the localization of OLZ within the rat. C) Localization of the MS/MS ion species m/z 256 that displays the distribution of the N-desmethyl metabolite of OLZ. D) MS/MS ion image of m/z 272, the 2-hydroxy methyl metabolite. Species were identified using the MS/MS capabilities of the TOF mass analyzer that breaks molecules into small component pieces and identifies them based on their fragmentation patterns. The species were identified during the experiment and not preselected or labeled prior to analysis. Scale bar = 1 cm Reproduced with permission from reference [22].

Current research into clinical biomarkers is also performed with IMS. In these studies, comparisons of readily available disease and normal tissue are analyzed via MALDI. Samples preserved with either formalin-fixed paraffin embedded (FFPE) protocols or fresh frozen tissue protocols are amenable to this type of analysis [29]. A
recent report detailed the use of MALDI IMS to examine synovial tissue of patients with rheumatoid arthritis and patients with osteoarthritis to identify relevant biomarkers [30]. A 2010 study by Grey and coworkers described MALDI IMS analysis of chick hearts. The authors record precise spatial localization of several \( m/z \) values from different anatomical regions of a heart slice [31]. This study defined a basic proteomic profile for a normal heart, focusing on proteins in the mass range of 3-30,000 \( m/z \) and reporting on values obtained between 3-15,000 \( m/z \). These results can be used in future research to compare models of heart defects and learn more about proteomic changes in the developing heart.

While lower mass molecular species (0-25000 \( m/z \)) are easily detected in IMS experiments, there is increased interest to extend analysis to higher molecular mass species, including proteins greater than 25000 \( m/z \). Figure 1.3 provides a demonstration by the Caprioli laboratory of MALDI imaging of higher molecular mass proteins in the mouse epididymis [32]. In this study, the molecular weight range was extended beyond 40,000 \( m/z \) and proteins up to 35,870 \( m/z \) were mapped.
Figure 1.3: Imaging Mass Spectrometry analysis of a mouse epididymis, a highly complex, compartmentalized organ with distinct areas containing different protein species that can be tracked via their distinct $m/z$ values. A) H&E stain of a tissue section after analysis with discrete regions annotated to the right of the picture. B) Averaged mass spectrum from all of the pixels from the tissue image. C) False color ion intensity maps with each color corresponding to $m/z$ values derived from different proteins in the compartments of the tissue. Reproduced with permission from reference [32].
1.3.2 Neuronal Samples

IMS has proven to be an especially valuable methodology to examine neurological samples. Brain tissue samples are ideally suited for mass spectrometric analysis as many of the key neurotransmitters are either peptides or small molecules. Beginning in 2000, Sweedler and coworkers have extensively mapped the neuropeptide distribution in the brain of the marine mollusk, *Aplysia californica* [33]. They have profiled samples ranging from intact ganglia [34], to cell cultures [23], to single cells [35], and the releasate from individual neurons [36, 37]. The Li laboratory has also performed extensive characterization of the neurological system of the crab, *Cancer borealis* [4, 38, 39].

1.3.3 Imaging of Pharmaceutical Compounds

Of particular interest to the readers of this journal are specific examples in which IMS has been used to track the distribution of drug molecules and drug metabolites as highlighted in Figure 1.3. The relatively rapid execution and chemically specific information provided by MALDI IMS experiments offers an enticing tool to enhance the pre-clinical trials of pharmaceutical compounds [13]. For a targeted approach, using the known molecular mass of a compound offers a distinct advantage in defining its localization allowing experimenters to easily distinguish the target from unknown species and/or electronic and chemical noise. Proteomic changes arising from treatment with a specific pharmaceutical compound can be tracked in conjunction with the distribution and metabolism of the compound.
Imaging of pharmaceutical compounds and their metabolites is quite challenging because of the small size of the drug molecules, but with high resolution mass analyzers these species can be successfully isolated in the lower mass regions (100-500 Da) [40]. In a 2007 study conducted by Drexler and coworkers MALDI IMS was used to examine fixed tissue samples of rats dosed with differing amounts of the prodrug BMS-X-P [41]. Using targeted IMS, the researchers identified the active form of the drug in various organs of the treated rats. They also used complementary techniques such as laser capture microdissection (LCM), MALDI MS and high pressure liquid chromatography (HPLC) MS to support their findings. Using what was known about the metabolism of the BMS-X-P, the spleens of dosed animals were examined via IMS for the full form of the drug compound as well as various metabolites. Once spatial distributions of the metabolized prodrug were generated, MS/MS sequencing was then used to confirm the identity of the compound using diagnostic fragment ions. This study represents an excellent example of a targeted IMS study, complemented with more traditional methods, that enables researchers to obtain biologically relevant information about drug metabolism of a specific compound in a spatially defined context.

In a study from 2012, Shanta and colleagues addressed the issue of matrix interference in low mass regions and drug/drug metabolite detection via MALDI IMS [42]. They detail the development of a binary matrix compound that has significantly reduced interference peaks from 0-500 Da. After verifying that the matrix did not interfere with drug detection they performed MALDI IMS on brains, kidneys and livers obtained from mice treated with donepezil, a frequently used therapeutic for the
treatment of Alzheimer’s disease. By comparing untreated controls to animals treated with varying doses, the researchers found the active form of the drug in both the kidneys and the brain of the mice but not in the liver. In addition to locating the drug based on its $m/z$ value, they performed MS/MS analysis to confirm that the appearance of $m/z$ 380 was due to the presence of the drug and not from some other source. The authors were able to successfully locate donepezil in multiple organs. The results also indicated that the drug is in its active form in the same region of the brain that commonly hosts β-amyloid plaques indicative of Alzheimer’s disease within 60 minutes of administration. This study demonstrates the use of IMS to obtain pertinent information about the localization of a drug molecule not readily observable by other means. Previous methods for establishing whether donepezil had reached its target organ were indirect, based on measurements of diagnostic peptides found in serum, but with the use of IMS the authors were able to directly observe the active drug molecule at specific sites in the brain corresponding to known sites of diseased tissue.

In 2011, the Castner group published a study that used Secondary Ion Mass Spectrometry (SIMS) IMS to probe HeLa cells treated with the anticancer drug bromodeoxyuridine (BrdU) [43]. The authors detected molecular and atomic species from BrdU inside the nuclei of the cells and performed depth profiling of individual cells. With a combination of Atomic Force Microscopy (AFM) and SIMS MS they were able to gain a better understanding of the 3D profile of individual cells. The IMS results demonstrated that BrdU was only present in the nuclei and not in the cytoplasm. The authors also evaluated the efficacy of various sample preparation protocols for SIMS
analysis by measuring membrane integrity as a function of intracellular contents found outside the cell membrane. This study demonstrates the ability of SIMS IMS to detect subcellular localization of multiple kinds of analytes. Furthermore, it demonstrates the pairing of two techniques with high spatial resolution, AFM and SIMS IMS, to give extraordinarily detailed information about individual cells.

For an excellent example of Desorption Electrospray Ionization (DESI) IMS of pharmaceutical compounds, a publication by the Cooks group describes the detection of the antipsychotic drug clozapine and its predominant metabolites in different tissues harvested from orally dosed rats [44]. The authors demonstrate rapid and accurate detection of the drug and its metabolites that correspond to previously known tissue distributions for the lungs, brains and kidneys of the animals. The authors also discovered a previously unknown localization of the drug in the testis. A full MS scan of the tissue via DESI was used to simultaneously detect not only the drug molecules but also many endogenous ions in each tissue. They then used well-established LC-MS methods to confirm their IMS findings. Quantification of drug levels in each tissue at select time points was done by comparing signal intensities of the drug ion to the phosphatidylcholine ion in three serial sections of tissue. These values were normalized to the highest signal intensities detected for each experiment and then plotted against LC-MS/MS profiles of drug detected in the plasma of the animals. They report an $R^2$ value of .9669 for the correlation of IMS values to LC-MS/MS values. This study highlights the flexibility of DESI IMS and demonstrates excellent potential for the utility of DESI IMS in the field of drug and drug metabolite detection. The combination of
target- and discovery-based approaches to IMS promise a bright future in drug research for IMS techniques.

1.4 Sample Preparation

Sample preparation is a critical part of the IMS experiment. The schematic in Figure 1.1 outlines the general steps in the IMS workflow. Depending on the specifics of the experimental design, each method of analysis has different requirements for sample preparation. In this section we will consider sample preparation protocols that are common to all IMS imaging. We will also consider those sample requirements that are unique to the specific types of IMS. There will be a special focus on the application of MALDI matrices in section 1.5.1.2 as the matrix application is currently the most critical determinant of spatial distribution, ionization efficiency and lateral spatial resolution (smallest achievable pixel size in the X/Y plane) in MALDI IMS. For additional information on sample preparation for IMS, please see the excellent recent review by Goodwin [17].

1.4.1 Sectioning of Biological Samples for IMS

Tissue preparation protocols for IMS, regardless of the ionization method, are primarily concerned with preserving the original spatial integrity of the tissue/biological specimen. Protocols also focus on how to maximize ionization of the class of molecule being analyzed and obtain spectra free of contamination and artifacts. Preventing analyte migration, degradation and contamination are the three main goals of sample preparation protocols for any kind of sample or type of IMS to be used.
Routine IMS protocols for the examination of tissue, tissue mimics and other three dimensional (3D) structures require the samples to be sliced into very thin sections. The most common way to prepare these samples is to section them on a cryostat with a temperature below -20°C [45]. Given that most samples are flash-frozen almost immediately upon collection, cryostats offer a relatively easy tool for precise and efficient sample sectioning. It should be noted that the Optimal Cutting Temperature (OCT) medium sometimes used in other cryosectioning applications is not suitable for IMS samples. OCT is a polymer and produces substantial background interference [46, 47].

Gelatin Assisted Sectioning (GAS) is an excellent substitute for OCT embedding [1, 38]. GAS is a relatively simple method in which fresh tissues are embedded in warm gelatin and then flash frozen immediately after the gelatin hardens. Gelatin is a liquid above 37°C and cools quickly, thus preventing changes to the sample due to excessive heat or exposure to the atmosphere. Additionally, gelatin is easy to handle and is tolerant of the freezing and cutting process. MALDI matrices do not crystallize on the surface of gelatin; therefore the gelatin itself does not ionize and does not impede ionization of analytes while still holding the tissue in its native state throughout the sample preparation and analysis [1, 48].

Slices of tissue that are 10-12 µm thick are suitable for both MALDI and SIMS workflows. Slightly thinner sections can be used for DESI experiments. Depending on the application a thicker section may be used, with an upper limit generally around 20 µm [49, 50]. Sample thickness is not a key determinant of ionization efficiency in SIMS or
DESI but can have a large impact for MALDI. In MALDI MS, a UV laser ionizes only the sample surrounded by the matrix molecules; therefore, enough material needs to be present to allow the matrix/solvent mixture to extract a sufficient amount of analyte for ionization. The amount of material available for ionization is a function of the sample thickness.

Following sectioning, the samples must be mounted on conductive plates or indium tin oxide (ITO)-coated slides for MALDI or SIMS analysis. Both of these approaches require the sample to be charged by applying a voltage to the sample plate (also referred to as a target) to allow for ionization. The most popular method for attaching the samples to the target is thaw mounting, in which the cold tissue is placed onto a room temperature target and the tissue adheres to the target as it thaws. Alternatively, cold tissue can be placed onto a cold target and both the sample and target plate brought to room temperature together [47]. Each method has advantages and disadvantages. For instance, there is evidence that the use of a room temperature slide or plate leads to a thin film of ice crystals that melt on the tissue surface and delocalize water-soluble proteins, leading to poorer spectra. However, the cold mounting method is very difficult to do with small or thin pieces of sample as they tend to crinkle or deform as they adhere to the cold plate [51].

Once the sample is mounted, a picture of the sample and target plate for co-registration is obtained. This step is especially important when dealing with tissues that have well-differentiated regions of biological relevance, as it will facilitate the alignment of images after the sample has been analyzed. The picture is also used as a reference
when setting up the imaging software, as it is often difficult to get a clear picture of the sample from the camera in the instrument, in the case of SIMS and MALDI. When using clear glass slides, marking the underside (which is non-conductive) with reflective ink helps with co-registration and keeping track of where the instrument is focused on the sample.

1.5 Ionization Approaches for Imaging

To date, MALDI IMS has proven to be the most robust and adaptable imaging technology but other ionization methods such as SIMS and DESI provide valuable information that complements other imaging techniques [52]. Each method discussed does have certain biases with regard to size and type of molecule most readily detected, but when combined with other forms of analysis each IMS method offers unique insight to a study. It should be kept in mind that there are additional ionization approaches, other than the three mentioned in this paper, that can be used either alone or in conjunction with other techniques, some examples include Atmospheric Pressure MALDI (APMALDI) [52], Laser Ablation Inductively Coupled Plasma Mass Spectrometry (LA ICP MS) [19] and Surface-Enhanced Laser Desorption/Ionization (SELDI) [19, 53]. Due to space limitations, we will only review the most popular IMS methods.

1.5.1 MALDI

MALDI MS employs the combination of high-energy laser pulses with an acidic matrix/solvent combination for the extraction, desorption and ionization of analytes from the surface of a sample [54]. The matrix molecules serve several purposes:
extracting analytes from the surface of the sample, surrounding and separating the analyte molecules and absorbing energy from a UV or IR laser pulse. Most instruments use either N₂ lasers with a 337 nm wavelength or frequency-tripled Nd:YAG lasers operating at 355 nm. Matrix molecules are optimized to absorb energy at these wavelengths to prevent premature decay or degradation of analyte molecules due to laser irradiation [55].

While the exact mechanism of ion formation in MALDI remains a subject of debate, the simplest description of the process is to imagine the laser irradiation causing an “explosion” at the surface of the sample underneath the beam. This explosion creates a plume of material in the gas phase just above the sample surface that contains a mixture of analyte and matrix ions as well as neutral species. The plume of desorbed material is then accelerated into the mass analyzer from the source with the application of an electric field. The role of the mass analyzer will be described further in section 1.6.

In order to account for ionization biases and differences in chemistry among classes of analytes, tailoring a method to a specific mass range is recommended. This means that while multiple types of molecules can be detected via MALDI IMS, it is most prudent to divide the analyses into experiments that examine small segments of a mass range. For example, typical mass ranges for analyses would be 0-500 Da for small molecules, 500-3000 Da for peptides and lipids and 8 kDa-50 kDa for proteins. Dividing up the mass range increases analyte sensitivity significantly as compared to scanning the entire mass range (0-50 kDa) in a single run. In order to perform imaging experiments, a MALDI mass spectrometer must have an auto-raster function as well as software...
capable of processing and visualizing the data. Most instrument manufacturers sell this software as a package.

1.5.1.1 Sample Preparation for MALDI Analysis

As neither SIMS nor DESI require further sample treatment, once the sample is in an acceptable form it can be subjected to IMS analysis. Samples for MALDI analysis require several additional factors to be considered prior to the IMS experiment. Once mounted to the target, the tissue is subjected to washing procedures to remove artifacts from the sectioning process, salts, and lipids (unless these are the desired targets of analysis). During these washing procedures, great care must be taken to avoid delocalization of the chemical species within the tissue. Multi-step washing procedures are routinely used when proteins and peptides are the desired targets of analysis [49]. There is also evidence that washing the samples can aid in matrix crystallization [45].

Washing procedures vary greatly depending on the analyte of interest for a given experiment. As mentioned above, MALDI IMS can be used to examine the spatial distribution of molecules spanning a range of masses, from relatively small drug metabolites to large, intact proteins. In light of this fact, it is important to choose a washing procedure that best enhances the target analyte signal while removing or suppressing interfering species. Most protein and peptide imaging protocols use cold mixtures of water and ethanol or isopropanol with an increasing concentration of alcohol over the course of a few washes [1, 3, 45, 51]. The use of cold solvents minimizes the solubility and delocalization of large molecules such as proteins and
peptides [45]. Normally sections are washed for 30 seconds to one minute and are placed in a desiccator immediately after the washing steps to remove residual water or solvent that could lead to analyte delocalization if left on the sample surface. It has been reported that adding low percentage organic solvents such as acetone, chloroform or acetic acid to the washes can significantly improve protein spectra without being detrimental to matrix application [56-61]. When imaging drugs and drug metabolites or other small molecules, it is generally better to do an abbreviated wash or not wash the sample at all as it is possible that most of the compounds could be washed away [40, 62].

1.5.1.2 Matrix Selection for MALDI Analysis

Once the sample has been sliced, affixed to the appropriate target plate and any necessary washing performed, the MALDI matrix must be applied. A molar excess of matrix is required for optimal ionization. The acidic matrix chemicals are used at saturating concentrations in solvents such as methanol, ethanol, isopropanol, acetonitrile, or any number of other volatile organic solvents. In general, most solvents and solvent combinations are fairly universal and can be used with most matrices. There are an abundance of matrix chemicals available for use and some are more compatible with certain classes of molecule than others. For example, dihydroxybenzoic acid (DHB) is more amenable for analysis of lipids and small peptides whereas sinapic acid (SA) is often the matrix of choice for analysis of large proteins [49]. Some matrices such as dihydroxyacetophenone (DHAP) and similar compounds may require additional
additives, such as heptafluorobutyric acid, to increase their stability when put under vacuum in the MALDI instrument [63]. Furthermore, commonly used matrices such as α-cyano-4-hydroxycinnamic acid (CHCA) cause interference in the lower mass regions due to some fragmentation of the matrix material itself (0-500 Da) and are often mixed with additives such as ammonium salts, citrates or phosphates in order to suppress or remove interfering signals and improve sensitivity [64]. These examples are provided to highlight the importance of choosing the proper matrix for the desired application; a list of several commonly used matrices and their applications is provided in Table 1.1 [45, 54, 65, 66]. The combination of matrix extraction of analytes and powerful laser ionization, MALDI offers the largest dynamic range of the ionization techniques discussed and provides the best opportunity to examine protein, large peptide and small molecule distributions.
TABLE 1.1
COMMONLY USED MALDI MATRICES, THEIR MOLECULAR MASSES, ABSORBANCE WAVELENGTHS AND APPLICATION

<table>
<thead>
<tr>
<th>Type of IMS</th>
<th>Ionization Source</th>
<th>Preferred Analytes</th>
<th>Mass Range</th>
<th>Spatial Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALDI</td>
<td>UV/IR Laser- soft ionization</td>
<td>Lipids, Peptides, Proteins, Small Molecules</td>
<td>0-50,000 Da</td>
<td>Tens of microns-millimeters</td>
</tr>
<tr>
<td>SIMS</td>
<td>Ion gun- hard ionization</td>
<td>Lipids, Small Peptides, Small Molecules</td>
<td>0-2000 Da</td>
<td>Sub-micron-Several hundred microns</td>
</tr>
<tr>
<td>DESI</td>
<td>Solvent spray- soft ionization</td>
<td>Lipids, Peptides, Small Molecules</td>
<td>0-2000 Da</td>
<td>Several hundred microns-millimeters</td>
</tr>
</tbody>
</table>

Adapted from 45,54,65,66

1.5.1.3 Matrix Application for MALDI Analysis

There are numerous ways to apply MALDI matrix for IMS analysis, some approaches are specialized for certain applications, such as peptide or large protein imaging, and others that are applicable to multiple types of analyses. Here we will explore several of the options available and discuss tailoring specific approaches to produce the desired experimental outcome.

Application of matrix solutions by hand via pipet is the quickest and easiest method. When applied to small samples however, this method can lead to widespread delocalization of extracted proteins and peptides because of the large volumes of solution compared to a relatively small sample area and inaccurate placement of matrix droplets. There are a variety of automatic sprayers, acoustical spotters and chemical
printers that can overcome the inaccuracies of hand spotting and can be adjusted to deposit very small volumes (pL-nL) of matrix solution in ordered arrays or patterns [49]. These devices are exceptionally useful for the imaging of whole organ slices as their reproducibility is excellent. Also, the patterns of matrix deposited are easily compatible with raster patterns in MALDI instruments. Gas nebulizers and airbrushes are also commonly used to apply a thin “film” of matrix to the sample surface. The nebulizer or airbrush is held far enough away from the sample that only a very fine mist of matrix solvent reaches the sample. Several passes across the surface of the sample with several seconds to dry in between passes are required. The end result is a fine coating of matrix crystals that is evenly distributed across the sample. This method does not leave the spaces that occur with the ordered patterns or arrays from chemical printers or acoustical spotters that makes it ideal for non-tissue samples. The airbrush application method requires a fair amount of skill to generate reproducible coverage but has been used to analyze multiple types of molecules from a variety of samples [41].

The size of the matrix crystals deposited on the tissue is a significant determining factor for spatial resolution. Under ideal experimental conditions, the laser spot size would determine the pixel size in the final MS image; however, this is often not the case. If the crystals are larger than the laser spot then the information acquired for each pixel will be mixed and not truly representative of the sample directly underneath the beam. This phenomenon is especially problematic when imaging mammalian cell culture samples with small diameter cells, as large matrix crystals will likely extract
proteins and/or peptides from multiple cells causing the distortion of important localization features.

Of the common methods of matrix application, the sublimation of matrix directly onto the sample affords the smallest crystal size with the least potential for delocalization of analytes [59]. Briefly, sublimation involves applying a dry coat of fine matrix crystals to a sample surface under high vacuum. After sufficient dry matrix coverage is achieved the sample is then “recrystallized” which involves exposing the matrix-covered sample to a solvent in a heated, airtight chamber. The recrystallization allows analyte extraction from the sample surface in a controlled manner without a washing or spraying motion that could delocalize analytes from their endogenous distributions [59]. Thus far, sublimation and recrystallization offers the best spatial resolution in terms of crystal size, which allows the laser spot of the MALDI to define the spatial resolution of the experiment [67]. The modes of matrix deposition discussed are mostly optimized for tissue applications but can be adapted for use with other samples like cell cultures. As mentioned briefly, the size difference between whole tissue slices and slices of much smaller in vitro tissue models represent the largest challenge for matrix application. The small size of these models also means that even slight movement in analytes from their native positions is amplified. Avoiding large volumes of matrix solvent is critical to maintaining optimal spatial resolution when performing IMS on small samples.
1.5.2 SIMS

While MALDI MS is an excellent technique to obtain information about large proteins and peptides, albeit at limited lateral spatial resolution, SIMS IMS is optimized for the ionization of much smaller analytes, such as lipids, metabolites and pharmaceuticals. A huge advantage of SIMS IMS is the higher spatial resolution compared to either MALDI or DESI imaging. Table 1.2 offers a brief overview of... Spatial resolution for SIMS IMS is routinely in the single micron range and smaller [52].

**TABLE 1.2**

**COMPARISON OF IMS APPROACHES DETAILING IONIZATION METHODS, MASS ANALYZERS, OPTIMAL ANALYTES, MASS RANGE AND RESOLUTION**

<table>
<thead>
<tr>
<th>Type of IMS</th>
<th>Ionization Source</th>
<th>Common Mass Analyzer</th>
<th>Optimal Analytes</th>
<th>Mass Range (Da)</th>
<th>Spatial Resolution (μm)</th>
</tr>
</thead>
<tbody>
<tr>
<td>MALDI*</td>
<td>UV/IR Laser Soft ionization</td>
<td>TOF</td>
<td>Lipids, Peptides, Proteins, Small Molecules</td>
<td>0-50,000</td>
<td>30-50</td>
</tr>
<tr>
<td>SIMS#</td>
<td>Ion gun Hard ionization</td>
<td>TOF Magnetic Sector Orbitrap</td>
<td>Lipids, Small Peptides, Small Molecules</td>
<td>0-2000</td>
<td>0.5-1</td>
</tr>
<tr>
<td>DESI$^\S$</td>
<td>Solvent spray Soft ionization</td>
<td>Orbitrap</td>
<td>Lipids, Peptides, Small Molecules</td>
<td>0-2000</td>
<td>100</td>
</tr>
</tbody>
</table>

*Denotes references 3,45,55
#Denotes references 19,68
$Denotes reference 77
SIMS instruments use ion guns to provide a beam of high-energy primary ions. In this context, the phrase “primary ions” is used to describe the ions produced by the ionization source for the purpose of generating analyte ions from the sample. Primary ions are not the ions detected; rather, they are used to generate secondary ions from the sample. The primary ion beam is placed at an angle to the surface of the sample and as the primary ions strike the sample, a plume of secondary ions is ejected into an electrostatic field that directs the ions into an adjacent mass analyzer [68]. The bombardment of the sample surface with the high energy primary ions causes extensive damage to the region closest to the initial impact, hence SIMS is known as a “hard ionization” method [19]. Because of the intense energy involved, the secondary ions that represent the surface of the sample are typically small fragments of the molecules and atoms not directly impacted by the primary ion beam. SIMS does not have the capability to ionize large analytes but is exceptionally useful at ionizing lipids and other organic molecules. The upper mass limit of SIMS experiments is around 2 kDa because there is no chemical extraction by matrix solvents and because of the extensive fragmentation of surface molecules, additionally molecular species deeper than a few nanometers are not detected [19, 47, 52].

There are two categories of SIMS imaging: static SIMS and dynamic SIMS. Static SIMS is normally used for biological samples because of its ability to detect relatively large molecules such as small peptides and lipids; however, dynamic SIMS has been shown to be an excellent tool for tracking drug distribution, especially at subcellular resolution [69].
1.5.2.1 Sample Preparation for SIMS Analysis

SIMS analysis requires ultrahigh vacuum conditions (<10e-6 mBar) to facilitate ion transfer after desorption from the sample surface, making analysis of biological samples difficult. While no sample prep is required for SIMS analysis, there are several optional techniques that can be performed depending on the desired application. One widely used sample preparation method is freeze-fracturing, which requires the specimen of interest to be grown or placed on some sort of conductive surface (such as silicon), sandwiched between a second piece of material and quickly frozen in liquid nitrogen. The two pieces of the sandwich are pulled apart to fracture the cells and allow subcellular measurements of ion concentrations [70]. While the method is highly sensitive and useful, it is not ideal for high-throughput analysis and requires considerable skill to obtain useful data.

Another popular method for the examination of biological samples via SIMS is freeze-drying, in which specimens are flash frozen in liquid propene or propane, stored in liquid nitrogen and slowly brought to room temperature in a vacuum to prevent condensation and to remove residual water in the sample [70, 71]. Alternatively, samples can be formalin cross-linked to chemically preserve the spatial distribution of molecules within a cell. Each of these methods has its pitfalls, but the end result is to preserve the sample in a biologically relevant state with spatially accurate distributions of analytes. The freeze-fracture method requires the extra step of using electron microscopy to examine the sample plates to determine the location of each cell and whether the cell was fractured along its apical or basal membrane [71].
The sample preparation approach depends on the type of SIMS being performed [72]. Active research exploring more amenable procedures is ongoing. Recently, peptides were detected from a tryptic digest of a tissue specimen using static SIMS. This approach yielded several relatively high molecular mass fragments with the assistance of a water matrix containing trifluoroacetic acid [73]. This and other method developments are propelling the adoption of SIMS imaging for more biological species.

1.5.2.2 Static SIMS

In static SIMS analysis, the “fluence” (number of primary ions) of the primary ion beam is pulsed and kept relatively low. This type of ion beam allows for detection of larger species but at much lower sensitivity, meaning the number of ions detected per pixel in the image is low [19]. The impact of the ions from the primary ion source causes damage to the sample surface by obliterating the sample directly underneath the beam. The impact initiates a chain reaction of kinetic energy that spreads within ~10 Å of the primary ion beam, ejecting small fragments of the sample that are ultimately detected. As a result, sampling the same surface area multiple times will only provide information from damaged surface. To be sure that each spectrum acquired is representative of undamaged sample, very low primary ion doses are used to damage less than 1% of the surface area [47]. This dose of primary ions is known as “the static SIMS limit” [47] or “the static limit” [19] and is what differentiates static SIMS from dynamic SIMS.

Static SIMS instruments have submicron spatial resolution, which is by far the highest resolution routinely achievable in IMS studies to date [74]. Typical ion sources
for static SIMS for biological applications are large clusters of ions such as gold clusters, SF5+, C60+ or Bi3+ as these tend to have relatively soft impacts. With the large size of the ion clusters, larger fragments are expelled from the samples thus accounting for the increased mass range of static SIMS [19, 68].

1.5.2.3 Dynamic SIMS

In dynamic SIMS, a higher energy, continuous primary ion beam is used. The beam erodes the surface of the sample and provides more ions per pixel. The secondary ions are also much smaller pieces than those generated by lower energy beams, usually just small organic fragments of larger molecules [68]. By measuring the isotopic distributions of monoatomic species such as Na+, Ca2+, Mg2+ and K+, inferences about the physiologic condition of cells, quantification of these species and even identification of individual organelles is possible [72]. Typically, dynamic SIMS images are co-registered with images from other high spatial resolution techniques such as AFM or Scanning Electron Microscopy prior to the SIMS experiments because of the destruction of the cells by the ion beam [43]. Due to its propensity to readily detect small organic molecules at isotopic resolution, dynamic SIMS is an excellent technique for examining drug distribution on a sub-cellular or even individual organelle scale [72].

1.5.3 DESI

DESI is a soft ionization technique that was only recently applied to IMS. As opposed to MALDI and SIMS, both of which require the sample to be in a vacuum chamber, DESI is performed in ambient conditions and requires no addition of matrix
chemical or further sample preparation [52]. DESI generates ions directly from the sample surface using a small jet of electrostatically charged solvent positioned at an angle to the sample surface, much like SIMS. A second collection tube is responsible for collecting the ions generated after the jet of solvent extracts them from the sample surface [52]. Since DESI is a soft ionization technique, unlike SIMS, there is very little fragmentation of the analytes due to the ionization energy, additionally, DESI does not require any sample preparation and does not ablate or damage the substrate being imaged [75].

The exact mechanism responsible for the formation of charged ions desorbed from the surface of the sample in DESI is still under debate. The simplest analogy ascribed to the mechanism thus far is “droplet pick-up” [75]. It is thought that ions are generated by an initial “sample wetting” step in which primary solvent molecules extract accessible molecules from the surface of the sample. These molecules are ejected from the sample surface by the impact of secondary solvent droplets. The secondary droplets react with the analyte molecules in the gas phase to produce charged analyte species that are subsequently swept into the collection tube leading to the mass spectrometer [76]. The ionization reactions occurring in the gas phase are the same mechanisms that occur in conventional ESI and therefore the spectra obtained through DESI are multiply charged and very similar to those obtained in an ESI experiment [52, 76].

Because DESI does not require sample preparation or a matrix, detection of the analyte of interest can be optimized most easily by using different solvent spray
combinations and varying the spray conditions. Spray parameters that can be adjusted include spray voltage, angle of incidence and tip distance from the sample. Addition of various chemicals to the spray solvent has a large impact on the gas-phase ionization reactions which, in turn, allow for further modification of the desired species analyzed in a given experiment [75]. DESI has been shown to be capable of detecting a wide range of analytes including drugs, polymers, proteins and lipids [2]. Tissue imaging applications for DESI include lipid imaging of different disease states, profiling of cancerous tissue and multi-modal imaging experiments that combine information from several different imaging techniques to facilitate an extremely thorough analysis of the sample [50, 77].

In work by the Cooks laboratory, mouse brain samples were imaged by DESI, MALDI and traditional histology (Figure 1.4) [50]. The rapid analysis time and the fact that various types of mass analyzers can be coupled to DESI ionization sources make it an attractive technique for tissue imaging of all kinds.
1.6 Instrumentation

The first MALDI imaging studies were done manually using a 20 Hz laser that required several hours to completely image an area roughly 1 mm$^2$. Within a few years of the first imaging experiments, automated imaging programs were developed to decrease sample acquisition time and compress file size [78]. A recent report described a modified instrument with a 5 kHz laser that can acquire 185-mm$^2$ area in less than 10 minutes [79]. Advances such as these offer rapid acquisition time, increased spatial
resolution and greater sensitivity. With continued improvements in all aspects of analyte desorption, ionization and detection, IMS promises to become an essential tool for proteomics research of all kinds. Below is a brief description of the types of mass analyzers commonly used in imaging research.

1.6.1 TOF Mass Analyzers

Time-of-Flight (TOF) mass analyzers are remarkably useful instruments for the analysis of biomolecules. They provide high mass resolution (1,000-25,000 FWHM), excellent mass accuracy, a rapid acquisition rate and large mass range [80]. The analyzer consists of a flight tube under vacuum with the ion source and focusing ion optics at one end and typically a micro-channel plate (MCP) detector (or multiple detectors for MS/MS sequencing) at the other. Multiple kinds of ionization sources can be used with a TOF analyzer, all of which rely on various types of pulsed ion techniques such as a rapid burst of UV laser light in the case of MALDI or a concentrated beam of primary ions for SIMS [7, 47]. The TOF instrument measures ions produced from the bombardment of the sample and calculates their mass based on the time it takes them to travel through a field-free drift zone (the flight tube). Because the initial kinetic energy of all the ions leaving the source is roughly equal, the time it takes the ions to reach the detector is proportional to their $m/z$, with heavier ions taking longer to reach the detector and smaller ions reaching the end of the flight tube more quickly. The mass analyzer converts the flight times into $m/z$ values based on measured flight times obtained from calibration of analytes with known masses [3]. TOF analyzers have a theoretically unlimited mass range with upper limits of analytes detected approaching 500 kDa [45].
1.6.2 Magnetic Sector Mass Analyzers

While TOF mass analyzers dominate the MALDI-IMS field, DESI and some SIMS sources can also be paired with either magnetic sector or Orbitrap mass analyzers. Magnetic sector instruments use strong magnetic and electric fields that are curved towards a detector so as to separate ions of different masses with the same kinetic energy. Ions of a preselected $m/z$ are detected with a resolving power of 105,000 [80]. The curved fields generated by powerful magnets cause ions that are heavier than the preselected $m/z$ value to fall “down” away from the detector and the ions that are lighter than the preselected value to be pushed “up” away from the detector. The drawback to the use of magnetic sector instruments is that they are not capable of scanning an entire $m/z$ range for all ions at once, instead, the $m/z$ values of interest must be predetermined and multiple scans must be performed to cover a large mass range. By placing multiple detectors after the sectors multiple $m/z$ values can be detected in a single pass [68].

1.6.3 Orbitrap Mass Analyzer

Orbitrap mass analyzers are a recent innovation in the MS community [81] and have only been commercially available since 2005; they provide a resolving power of 150,000 with a mass accuracy of 2-5 ppm and a wide dynamic range [82]. Orbitraps operate by trapping ions in an oscillating electric field generated by a central electrode kept at a constant voltage with outer electrodes at varying voltages [83]. Ions are introduced into the trap in discrete packets and then are separated by their different oscillations in the generated electric field. The ions are detected by determining the
frequency with which they oscillate around the central electrode as measured by the charges they produce in the outer electrodes. These frequencies are translated into readable $m/z$ values by fast Fourier transform and are proportional to $1/m/z$ of each ion in the packet introduced into the trap [80].

Each type of mass analyzer has its own strengths and weaknesses and the use of each should be varied according to the desired application. It should also be noted that “typical” pairings are referred to in this manuscript, but MALDI, DESI and SIMS can be paired with any number of mass analyzers, with the most popular pairings are MALDI-TOF, DESI-Orbitrap and SIMS-TOF or SIMS-magnetic sector (Table 1.2).

1.7 Data Analysis

Data analysis is a key aspect of IMS and must be done carefully in order to avoid misinterpretation of results and to gain useful information from an experiment. There are multiple factors that affect data analysis, which depend on optimization and careful execution of the aforementioned experimental steps to ensure that the integrity of the experimental data is maintained throughout the data acquisition process. IMS analyses have several steps that must be carried out in a timely manner to minimize sample degradation and day-to-day variations that can have a negative impact on the data. It is strongly recommended that every step in both the data acquisition and data processing workflows be optimized prior to combining all aspects into one contiguous experiment. As with most experiments that deal with large datasets, small inconsistencies in one portion of the workflow multiply into large errors [15].
Once each step of the process has been optimized and a standard operating protocol established, the combination of all of the steps of the actual IMS experiment give large lists of mass spectra that must be processed and interpreted the same way. The data processing must be finely tuned to best distinguish real analyte signals from background and chemical noise. There are a variety of data processing algorithms available, the details of which will not be enumerated here; the reader is instead referred to the excellent review by Jones et al. on IMS statistical analysis [15].

1.7.1 Statistical Analysis

There are numerous ways to process the spectra from an imaging experiment, whether the processing algorithms available with commercial software such as Bruker’s FlexAnalysis are used, or if the raw spectra are imported into another program such as MatLab (Mathworks, Natick, MA), the goal is the same: to obtain spectra that have been baseline subtracted and smoothed to give the best representation of the analytes detected [84, 85]. This is accomplished by smoothing and denoising algorithms that employ signal-to-noise ratios, user-selected intensity thresholds, and other complex statistical operations [15]. Once the spectra have been processed they are correlated back to their Cartesian coordinates on a representative optical image of the sample to generate a heat map, or ion intensity map, showing the distribution of each m/z species detected during the IMS run. The final step in data processing is the option to stack serially sectioned pieces of a 3D sample together to allow for analysis of all the images as a whole, offering an examination of analyte distributions in the context of a 3D structure such as a brain, kidney, liver or even 3D cell/tissue culture model [48, 86-88].
In addition to the 3D modeling, which will be discussed in more detail below, there are numerous other statistical analyses that can be performed on IMS datasets that compare the spectra among each other rather than refine the peaks in individual spectra as in the data smoothing. These methods revolve around determining significant changes between peaks found in all or some of the spectra. In general, these methods find peaks that are the most statistically significant with respect to ion intensity. These peaks are “clustered” together as the most intense. Further iterations of peak clustering based on decreasing intensity are performed until there are no more peaks above a user-defined intensity. New ion heat maps are drawn based on the difference in intensity between the clustered peaks. Clustering enables more accurate and useful interpretation of results by highlighting statistically relevant differences across the data set as whole and removing random fluctuations in signals based on chemical or electronic noise. Some examples of clustering methods include hierarchical clustering, multivariate analysis, principal component analysis (PCA) and a variety of biomarker discovery methods that allow for accurate and informative comparisons between the acquired mass spectra [15, 89-91]. Peak picking and alignment processes are further described in these sources [92, 93].

It is most prudent to use a combination of statistical tools. One common approach that has been suggested is to use PCA to first reduce the amount of data available and from this refined data set apply other forms of analyses to search for patterns among molecular species detected and potential biomarkers [94]. The chosen form of data analysis depends on the specific application, whether that is biomarker
discovery, tissue classification in correlation with histological grade, or identification of
disease versus normal tissue beyond the bounds of conventional histology [15, 84].

1.7.2 3D Reconstruction

The recent and ongoing development of 3D IMS modeling represents huge potential for the further development and refinement of IMS. Compiling 3D imaging datasets offers the opportunity to examine whole organ or even whole body proteomic analysis in small mammal experimental models [93]. 3D IMS consists of creating a “stack” of IMS images from the analysis of successive slices of a sample, typically a mammalian organ, that provide a comprehensive view of the entire structure in the X, Y and Z dimensions. Most of the discussion in this section is centered on handling and processing of data from multiple IMS experiments and therefore more computational in nature than experimental. There are several practical considerations when constructing a 3D model and different possible approaches for the compilation of 2D data into 3D datasets.

The first consideration for 3D imaging is alignment of images from different modes of imaging. The IMS ion maps can be combined with optical images of various kinds, including bright field, fluorescent images, or H&E stains. Another strategy is the correlation of IMS, optical and in vivo magnetic resonance (MRI, PET, CT, etc.) images into a single 3D volume [48, 86, 87]. Alignment of images from the different modalities is crucial and must be performed carefully in order to avoid warping or misalignment in the resulting 3D stack; this feat is accomplished using preexisting images from an
anatomy atlas or using well-defined architectural features discernible from optical images of each section of tissue [48, 88].

When applying 3D IMS to other structures with no defined architecture such as a tumor, another approach must be taken. The most straightforward method is the use of common points of reference readily visible in each modality; these points are also called “fiducials” or “fiducial markers” and can consist of something as simple as toner ink from an inkjet printer or a dye with a distinct molecular signature that is easily distinguished by MS. Cresyl violet and Ponceau S are commonly used [88, 95]. These points are placed on the conductive slide adjacent to the sample and must be placed in the same position relative to the sample on each successive slide. The advantage to using a colored dye such as Ponceau S is that it has a very distinct molecular signature, with an obvious red color that is distinguishable in both optical and MS methods and gives a fluorescent signature for use with fluorescent microscopy.

The second consideration of generating 3D data sets require the compilation of many 2D slices that inevitably require several days or weeks of data acquisition due to the labor-intensive workflow involved in imaging even a single 2D slice. As a result, there will be fluctuations between the spectra of images acquired on different days. As mentioned previously, even small differences will be compounded into larger errors in the form of misalignment in the 3D stack. To prevent errors, using a dye with a distinct molecular signature allows for visual alignment and provides an internal standard for normalization of the mass spectra, which helps minimize slice-to-slice variation when compiling the 3D image [95]. Since the dye’s molecular signature will include the same
peaks for every 2D slice, normalization to the slice with the most intense peaks from the
dye will help account for unavoidable variations amongst the slices [95].

Once a proper alignment method is selected, great care must be taken during
the sample sectioning process to keep the samples in the same orientation and
minimize creasing, folding and tearing of the thin sections [48]. Maintaining the same
orientation for each slice of sample is critical to maintaining spatial resolution of both
the X and Y coordinates of the IMS pixels in the individual sections and Z coordinates of
the resulting 3D stack. For example, if a 100 µm pixel is to be used in a MALDI IMS
image, then the sections to be imaged should be sliced 100 µm apart during the
sectioning so as to minimize the effects of section-to-section variation and provide
optimal resolution in the Z plane [48].

Finally, a critical step in compiling 3D IMS images is data reduction and data
processing. As previously discussed, IMS datasets are enormous and require substantial
computing power to process. To make the computing requirements more manageable,
spectral processing is required before the 2D images are compiled into a 3D stack [92].
This involves applying the clustering methods discussed in the previous section and
applying them to the 2D slices in the context of a 3D volume. This is different from the
previous discussion of clustering all of the spectra in an individual 2D slice but the same
clustering operations are employed. It is easiest to think of the summed, processed
spectra from each 2D slice as individual pixels in the 3D volume that will be further
processed and clustered in the same way as the original pixels in the 2D slices. The
purpose of further clustering the 2D slices in a 3D volume is primarily to account for
differences among slices and minimize variation resulting from experimental or instrumental sources so that the biological variation can be highlighted and examined [96]. The clustering method chosen must not compromise mass accuracy of the aligned peaks and achieve optimal resolution for highlighting spatial distribution of molecular species. There are also open source programs (BioMap, MatLab and ImageJ, among others) and commercially available software packages (FlexImaging, Bruker Daltonics, Bremen, Germany and SCiLIS Lab Software, Steinbeis Innovation Center for Scientific Computing in Life Science, Bremen, Germany) that aid in 3D image generation.

1.8 Challenges

1.8.1 Data Analysis

One crucial challenge in IMS experiments is the development of a standard, reproducible method of analysis that not only allows for the extraction of useful information out of each individual dataset, but that can also be applied to multiple datasets and agrees with other imaging methodologies such as histology [15, 89, 94]. Additionally, steps are being taken by several groups to reduce the amount of human interaction during data processing steps as this can introduce a large degree of subjectivity, especially in the area of biomarker or disease state research [84, 96, 97]. Automation of the spectral processing and correlation steps will not only increase reproducibility of results, but will also reduce the amount of time needed to perform an experiment. Automating throughput will advance IMS as a viable clinical method for
examining disease states, as it will only require researchers to prepare their samples and interpret the final results removing any subjective bias that might be involved.

The sheer size of the raw, unprocessed spectra from even relatively small sample images adds up and eventually presents a problem in the form of computing limitations. As improvements are made in instrumentation and sample preparation, the amount of useful signals acquired during an IMS experiment promises to increase [93]. While richer data sets have obvious advantages, limitations in available computing power and storage space may necessitate preserving only the relevant pieces of data (meaning actual changes to the sample as a result of treatment as opposed to fluctuations arising from chemical or electronic interference) from a large dataset for further analysis [84, 91, 96].

1.8.2 Spatial Resolution

Spatial resolution is another significant challenge in IMS analysis, in particular for MALDI and DESI. Laser spot size and analyte spreading due to the application of the matrix and tissue washing with solvent are the three primary limiting factors for MALDI imaging. Single cell resolution, that is the ablation and detection of molecules from one cell per laser shot, is not currently possible in routine MALDI IMS applications. The typical lateral spatial resolution of a commercially available MALDI instrument is 30-50 µm, while the typical diameter of a mammalian cell is 5-20µm [98]. Lateral spatial resolution of 20 µm can be achieved on some newer commercial instruments with special focusing optics but laser spots of around 50 µm are routine on older, non-modified commercial instruments [7]. With a laser spot of 50 µm, any features smaller
than 50 µm are incorporated into one pixel, meaning that important boundaries or small but important changes in localization that are smaller than 50 µm would be skewed or lost. Recently, Lagarrigue and colleagues imaged rat testes via MALDI IMS with a modified instrument capable of achieving a 20 µm spot size. Shown in Figure 1.5 is a high resolution image of rat testes which was chosen because of the highly compartmentalized tissue, consisting of closely arranged features that require precise mapping to reconstruct. Researchers were able to visualize IMS images of protein species between 2-16 kDa with 20 µm resolution, matching the IMS images with H&E staining that was performed on the same tissue sections [99]. This demonstration shows that MALDI IMS has the capability to achieve much better resolution than is currently standard in the field.

Figure 1.5: High resolution MALDI IMS of rat testes performed with a 20 µm laser raster. A) H&E stain of two compartments of the tissue denoted by the Roman numerals in the image. B) MALDI false color ion image for four different protein species precisely detected in each compartment. This experiment demonstrates a large step towards improved lateral spatial resolution in MALDI IMS experiments. Scale bar is 200 µm. Adapted with permission from [99].
Bouschen and co-workers reported a resolution of 2 µm on a modified instrument was used to examine analytes on a sub-cellular level in a renal carcinoma sample [67]. This again demonstrates the potential for MALDI IMS to approach the resolution achievable for SIMS IMS experiments. For the purposes of this review, only non-modified commercially available instruments are considered, however, the reader should be aware that a number of groups are making enormous strides towards single cell and sub-cellular resolution in MALDI IMS workflows, and should consult these excellent sources for further reading [19, 100].

While laser spot size is a critical factor in determining the lateral resolution of MALDI IMS experiments, there are other instrumental factors that must be considered in order to move towards single-cell resolution. When adjusting spot size, laser fluence should also be considered. The laser fluence is the amount of energy required before the sample will successfully desorb from the sample surface and ionize and fluence requirements vary as a function of spot size. A smaller laser spot requires a higher fluence and vice versa [55]. The energy distribution of the laser beam spot in many commercial instruments follows a Gaussian distribution with the highest energy area focused in the center of the spot, decreasing with increasing distance from the center [101]. As a result, it is possible to entirely ablate the region directly in the center of the spot without acquiring a useful signal. Just the opposite, areas in the outer regions of a large laser spot size may not receive sufficient energy for efficient ionization [47].

While newer instruments are equipped with next generation lasers that do not suffer as readily from an uneven energy distribution, approaches to control for uneven
laser fluence have been developed. Traditionally, features as large as or larger than the smallest achievable laser spot diameter were considered the definition of lateral spatial resolution assuming proper matrix coverage, however, Sweedler and coworkers developed a technique called “oversampling with complete sample ablation” to allow for imaging of features smaller than the diameter of the laser beam [98]. In this approach, the sample plate is rastered in increments smaller than the laser spot diameter and all the matrix is ablated from one spot before the plate is moved. As a result, analytes are only sampled from the incremental raster length, which can be substantially smaller than the laser spot size. This technique addresses the uneven ionization energy by exposing the entire sample to the same portion of the laser beam, which has the optimal conditions for efficient ionization. It also improves spatial resolution significantly without modifications to the instrument.

Similar to MALDI imaging, DESI imaging also has poorer spatial resolution when compared to SIMS. Spatial resolution in the first DESI imaging experiments was 0.5-1 mm [102]. Since then, progress has been made with spatial resolution reaching 40 µm on a normal, non-modified instrument [102]. More recently, 12 µm resolution was demonstrated with a new adaption of DESI, termed nano-DESI, which promises further improvements for even greater spatially resolved images [77].

Another limitation to DESI is a consequence of one of its advantages: due to the lack of an enclosed and vacuum-sealed chamber, contaminants from the atmosphere surrounding the tissue can be incorporated into the spray. This feature also presents the problem of ionization efficiency. DESI does not suffer as much from ionization efficiency
as SIMS, but improvements to the ion collection/transfer tubes are necessary to gain the maximum information from each sample and this is an area of active research for expanding the realm of DESI imaging [52, 75].

As discussed above, SIMS has the lowest spatial resolution (Table 1.2). Achieving sub-cellular resolution, quantification of ions from single cells and even identification of individual organelles have all been reported and discussed previously in this review. It should also be mentioned that these advantages to SIMS IMS analysis unfortunately preclude the practicality of using the technique to examine whole organs or even large cell culture samples. The amount of time required to perform the analysis of such large areas coupled with the massive amounts of data that would be generated make these types of analyses impractical. SIMS IMS is therefore most useful when used as a corollary technique to gain information about specific areas in a larger sample volume [103, 104].

Spatial resolution limitations also apply to the data analysis portion of the experiment when comparing or correlating ion heat maps to optical microscopy images with high spatial resolution. While useful data can be gained from a combination of these images, steps are being taken to provide high spatial resolution images even when the IMS technique itself is still limited in terms of spatial resolution. As discussed above, when spectra are processed they often undergo several clustering operations that serve to both reduce the physical size of the dataset and highlight relevant signals that correspond to actual molecular species detected as opposed to chemical or electronic noise. During the clustering process there is often a reduction in the resolution of the
resulting heat maps that can lead to important boundaries defined by detection of
different molecular species to be ill-defined or lost. Alexandrov and colleagues propose
an elegant solution to this problem by performing super-resolution segmentation of the
data, which improves the spatial resolution of IMS ion images without further dividing
already clustered peaks [105]. This example serves to illustrate the critical importance of
both data processing and attention to detail in IMS experiments.

1.9 A New Application: Cell cultures

While IMS has primarily been applied to harvested tissue samples, there is
growing interest in expanding its utility to other model systems. In particular, the
remainder of this review will focus on the new application of profiling cell culture
systems by IMS. Cell culture has long been a popular method to ask basic questions
about cell biology, biochemistry and cell structure. Compared to animal models, cell
cultures are relatively low cost, easy to manipulate and enable a wide range of assays
and experiments.

3D cell cultures are an especially attractive option to mimic in vivo tissues and
organs [51, 106, 107]. For example, the MCF10A cell line forms hollow acinar structures
that mimic the duct and lumen structure of human breast tissue [106]. The cell models
are extensively used to study normal breast development and the spatial distribution of
proteins in these structures has typically only been explored via immunohistochemistry.
The human colon carcinoma cell line, HCT 116, forms multicellular aggregates. These
multicellular aggregates are often called spheroids and develop oxygen and nutrient
gradients similar to those found in primary colon tumors, providing relevant information regarding *in vivo* molecular interactions in an in vitro model [108]. Models for 3D cardiac tissues are another realm of the ever-expanding 3D cell and tissue culture repertoire. Desroches and colleagues describe a system using cardiac myocytes and cardiac fibroblasts in coculture to produce “scaffold-free” heart tissue mimics which not only developed physiologically relevant structures in terms of developing cell-cell contacts but were also capable of spontaneously beating and could generate action potentials similar to those measured in freshly isolated heart tissue [109]. These models could be readily adapted to IMS protocols opening up new opportunities for proteomic heart disease research.

3D cell culture has expanded the number of experiments that can be performed in vitro [106, 110]. Compared to animal models, information can be gained from cell culture in a quick and efficient manner [111]. Numerous strategies have been developed to modify 3D cell cultures so that they model different organ systems. For example, when cell cultures are grown in the presence of purified basement membrane proteins, which mimic the extracellular matrix seen in tissues in the body rather than the plastic of cell culture dishes, the 3D structures increase in size [112]. The development of co-culturing models adds heterogeneity to model systems allowing for different cell types to interact just as they would *in vivo* [113]. Bioreactor technology serves to increase the size and complexity attained in a relatively short amount of time by growing the cells under conditions of shear stress that closely mimic the environment seen in blood vessels [114]. As researchers expand cell culture systems to include stem cells and
growth of multilayered tissue, these models promise to deliver increasingly accurate
depictions of the *in vivo* environment [115].

1.9.1 Imaging Cell Cultures via MALDI

MALDI imaging of cell cultures was first demonstrated by the Sweedler group [23]. Two-dimensional neuronal cultures from *Aplysia californica* were grown on glass beads embedded in parafilm. Using the stretched sample technique, the parafilm, and therefore, the cultures were stretched evenly in two dimensions [116]. The sparse neuronal cell cultures were parsed into individual cells, and potentially subcellular units, for analysis. Strong MS and MS/MS signals were detected corresponding to the pedal peptide, a neuropeptide that is abundantly produced in the pedal ganglia, demonstrating the ability to identify proteins via IMS methods.

Our laboratory has adapted many of the sample preparation protocols from tissues to the IMS analysis of 3D cell cultures [1]. Colon adenocarcinoma spheroids were grown to ~1 mm in diameter according to standard 3D cell culture protocols [117]. After 10 days of growth, spheroids were gently aspirated from their growing vessels via serological pipet and transferred to 24 well dishes containing a basement layer of solidified gelatin. Up to four spheroids can be transferred to a single well of a 24 well plate. Liquid gelatin at 37 °C is added to cover the spheroids and allowed to cool to room temperature. At room temperature, the solidified gelatin mass containing the spheroids is scooped out of the plate and frozen. The gelatin mass is then sectioned on a cryostat in the same manner as any tissue sample. We find that four slices from one block of embedded spheroids easily fit on a single ITO slide. As a result, either 16
different spheroids, or four vertical layers of one set of four spheroids can be examined on a single ITO slide. With this methodology, we have mapped differential distribution of analytes in HCT116 spheroids shown in Figure 1.6. Using a combination of MALDI imaging and LC-MS/MS we have confirmed spatial localization of cytochrome C and Histone H4 in three dimensions.
Figure 1.6: Depiction of 4 protein species detected on a slice from a tumor spheroid model grown from the colorectal carcinoma cell line HCT 116. Each species is detected in a spatially defined region of the spheroid and m/z value 12,828 is clearly localized to the inner portion of the spheroid slice. Detection of this species highlights the biologically relevant gradients recapitulated in 3D cell culture models by demonstrating proteomic differences between the inner necrotic core and other regions of the tumor spheroid mimicking tumor formation in vivo. Reproduced with permission from reference [1].
Special care should be taken when washing 3D cell culture samples for IMS as these models have no connective tissue or extra-cellular matrix beyond cell-cell contacts that helps hold animal tissue together. While many of the previously discussed solvents have been tested on 3D cell culture models, the wash protocols are often abbreviated or amended to account for the more fragile nature of these samples (Unpublished data, Hummon laboratory). After the washing steps are complete, storage at room temperature in a desiccator until the next step is usually sufficient to preserve the sample integrity, though storage for more than a few days is not recommended [58]. The same considerations for matrix application to small tissue samples also applies to 3D cell culture systems with the additional caveat that because of the small size of the cultures, typical automated matrix application devices may not provide the proper density or volume of matrix solution to allow for efficient ionization.

1.9.2 SIMS Imaging of Cell Cultures

SIMS imaging has also been used to examine cell cultures. In a recent study by the Ratner group, static SIMS imaging was employed to distinguish mammalian cell types [69]. Primary rat esophageal epithelial cells (REEC) were cultured with NIH 3T3 mouse fibroblasts. The cells were cultured on tissue culture polystyrene and imaged with a C60+ beam. Phase contrast images of the individual cell types and mixed cultures are shown in Figure 1.7a. SIMS images are shown in Figure 1.7b. PCA and partial least-squares discriminant analysis (PLS-DA) were used to identify the peaks that correlated with the two cell types. With the mass spectral data, the authors were able to identify the individual cell types in the culture. The authors state that this promising
development could be used to spatially locate cell types within a complex mixture or to examine the phenotypic purity of heterogeneous cellular populations.

Figure 1.7: An example of high resolution SIMS IMS on multiple cell culture substrates. A through C are phase contrast images of the cells in culture prior to SIMS imaging and D through F are the corresponding total ion maps from SIMS analysis. A and D are rat esophageal cells, B and E are NIH 3T3 fibroblast cells and C and F are a heterogeneous culture consisting of both rat esophageal and 3T3 cells. This figure demonstrates both the application of coculturing to provide a more accurate cell culture model of the in vivo environment and the capability of SIMS IMS to provide extraordinarily high-resolution images. It also demonstrates the ability of SIMS to identify different cell types in the same cell culture substrate. Scale bar is 100 µm. Reproduced with permission from reference [69].

1.10 Perspectives

IMS has made remarkable advances in the realm of imaging intact tissue samples. The aforementioned advantages of this technique (preservation of spatial resolution and obtaining multiple analytes in one experiment) can be combined with the growing field of 3D cell culture and tissue engineering to provide quick, inexpensive and
detailed information on disease states, drug delivery and drug metabolism on either a global proteomic or individual protein scale. Further development of de novo sequencing on these tissue mimics will only enhance the information this technique can provide. Additionally, the combination of IMS and cytochemical/ histological techniques will increase the knowledge gained from each experiment.

1.11 Acknowledgments

The authors thank the Notre Dame Mass Spectrometry and Proteomics Facility and especially Dr. Michelle Joyce and Dr. William Boggess for many helpful discussions. We acknowledge Leigh Weston for her assistance with Figure 1.1. Funding was provided by the 2011 Starter Grant from the Society for Analytical Chemists of Pittsburgh and the University of Notre Dame.

1.12 References


spectrometric and optical tissue imaging, Analytical Chemistry, 84 (2012) 1817-1823.


CHAPTER 2:
SINGLE CELL METABOLIC PROFILING OF TUMOR MIMICS

This chapter describes metabolic differences in the three biologically relevant regions of CRC tumor mimics grown from the HCT 116 cell line and was performed in collaboration with Dr. Richard Keithley who provided fluorescently labeled glycosphingolipids and performed the statistical analysis of the CE-LIF data as well as the cell size measurements. Undergraduate researchers Mark Metzinger and Andrea Rosado from Dr. Norman Dovichi’s lab performed significant portions of the CE-LIF experiments. We performed the cell culture and serial trypsinization experiments and analysis of the biological significance of the data. This study was published as an “Editor’s Choice” manuscript in Analytical Chemistry in 2013.

2.1 Abstract

Chemical cytometry employs modern analytical methods to study the differences in composition between single cells to better understand development, cellular differentiation, and disease. Metabolic cytometry is a form of chemical cytometry wherein cells are incubated with and allowed to metabolize fluorescently labeled small molecules. Capillary electrophoresis with laser-induced fluorescence detection is then used to characterize the extent of metabolism at the single cell level.
To date, all metabolic cytometry experiments have used conventional two-dimensional cell cultures. HCT 116 spheroids are a three-dimensional cell culture system, morphologically and phenotypically similar to tumors. Here, intact HCT 116 multicellular spheroids were simultaneously incubated with three fluorescently labeled glycosphingolipid substrates, GM3-BODIPY-FL, GM1-BODIPY-TMR, and lactosylceramide-BODIPY-650/665. These substrates are spectrally distinct, and their use allows the simultaneous probing of metabolism at three different points in the glycolipid metabolic cascade. Beginning with intact spheroids, a serial trypsinization and trituration procedure was used to isolate single cells from spatially distinct regions of the spheroid. Cells from the distinct regions showed unique metabolic patterns. Treatment with the lysosomal inhibitor and potential chemotherapeutic chloroquine consistently decreased catabolism for all substrates. Nearly 200 cells were taken for analysis. Principal component analysis with a multivariate measure of precision was used to quantify cell-to-cell variability in glycosphingolipid metabolism as a function of cellular localization and chloroquine treatment. While cells from different regions exhibited differences in metabolism, the heterogeneity in metabolism did not differ significantly across the experimental conditions.

2.2 Introduction

Three-dimensional cell culture is a relatively new in vitro system that recapitulates portions of the in vivo environment while retaining the benefits of traditional cell and tissue culture [1]. For some cell types, three-dimensional cell culture
can produce a multi-cellular tumor spheroid (MCTS) architecture. Like cells in tissue, cells in MCTS adopt polarity that can produce distinct spheroid microenvironments [2,3].

As tumors develop *in vivo*, cellular areas distal to blood vessels are deprived of nutrients and oxygen and cannot remove waste products from adjacent cells; nutrient deprivation leads to formation of hypoxic and necrotic areas [3-4]. Conversely, the parts of the tumor that are proximal to blood vessels are well-oxygenated and exposed to a variety of different growth factors secreted by neighboring cells; these nutrient rich regions are highly active and proliferative.

Three-dimensional cell culture mimics these gradients in vitro. Under proper conditions, HCT 116 colon carcinoma cells can aggregate into a single MCTS that reach up to 1 mm in diameter [5]. As the MCTS grow, the cells near the center become nutrient starved and ultimately die as the surrounding cells restrict access to the growth media. The cells on the outside of the spheroids remain metabolically active and proliferative [6]. The cells in the middle layers lapse into a quiescent state where they are metabolically active but not proliferating. Areas of necrosis and apoptosis increase with time, producing a central mass of cell debris surrounded by a layer of nutrient-starved but still living cells that are likely in a pre-necrotic state [7]. Serial trypsinization is a well-established technique to probe the microenvironments of MCTS [8-11]. Dilute trypsin is used to gently peel intact layers of cells from the outer regions spheroids. The process is repeated until the necrotic core containing only dead cells and cellular debris is reached.
Metabolic cytometry combines capillary electrophoresis with laser-induced fluorescence detection (CE-LIF) to probe metabolism in single cells [12]. In this approach, cells are incubated with a fluorescently labeled substrate. Single cells are then aspirated into a capillary and lysed. The fluorescent metabolites are separated and detected with CE-LIF.

In this paper, we use metabolic cytometry to investigate the metabolism of glycosphingolipids (GSLs) in different regions within MCTSs. GSLs constitute 5 - 20% of cell surface lipids, depending on the cell type [13]. GSLs are involved in a variety of cellular processes, and defects in their metabolism are associated with a number of diseases, including cancer [14-16]. These amphiphilic molecules consist of a nonpolar ceramide lipid tail and a polysaccharide headgroup. Figure A.1.1 (Appendix A) presents simplified overview of GSL metabolism in colon carcinomas [17-19]. Synthetic GSLs have been prepared with fluorophores attached to the ceramide tail for use as substrates in metabolic cytometry studies of several systems [20-29].

There is interest in using drugs to perturb metabolism in chemical cytometry. Chloroquine, an anti-malarial, inhibits catabolism in several ways: equilibrative transport brings it to the lysosome where it increases the intralysosomal pH from 4.7 to 6.2, it targets several lysosomal hydrolases, and the drug effects catabolism on a transport level, inhibiting cell membrane recycling and endosomal fusion with the lysosome [30-34]. Finally, chloroquine also acts an autophagy inhibitor and a novel potential chemotherapeutic [35-36].
2.3 Experimental

2.3.1 Chemicals and Reagents

TRIS buffer, sodium dodecylsulfate, agarose, ethanol, and chloroquine (diphosphate salt) were from Sigma Aldrich. CHES buffer, α-cyclodextrin, and paraformaldehyde were from Alfa Aesar. Glycine was from VWR. HCT 116 cells were obtained from American Type Cell Culture. Phosphate-buffered saline (PBS), trypsin, McCoy’s 5A cell culture media, and L-glutamine were from Life Technologies. Fetal bovine serum was from ThermoFisher. All buffers and stock solutions were prepared using distilled deionized water. GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 were kind gifts from Ole Hindgaul and Monica Palcic, synthesized as described previously [21-22].

2.3.2 HCT 116 MCTS Culture and Lipid Incubation

Aliquots of 6 nmol each from stock solutions of GM3-BODIPY-FL (in dH2O), GM1-BODIPY-TMR (in dH2O), and LacCer-BODIPY-650/665 (in ethanol) were combined with 9 nmol methyl-β-cyclodextrin (in dH2O) and dried using a Savant SpeedVac. Methyl-β-cyclodextrin acts as a lipid carrier, increasing cellular uptake of the fluorescent GSLs. Approximately 75 nmol of chloroquine (in dH2O) was added to another set of 6 nmol each fluorescent GSLs and this solution was also dried using the SpeedVac.

Three-dimensional HCT 116 MCTSs were cultured in McCoy’s 5A culture medium supplemented with 10% fetal bovine serum and 2.5 mM L-glutamine [5]. Approximately 6000 cells were seeded into each of 30 wells of a 96 well plate, resting on an agarose
meniscus. After 14 days in culture, the cells formed uniform spheres of ~1 mm diameter in the center of each well.

On the day of MCTS incubation, aliquots of lipids with and without chloroquine were each resolubilized in 15 µL ethanol and diluted to a final volume of 3 mL with culture medium containing only 5% fetal bovine serum. This dilution yielded final concentrations of 2 µM for each of the fluorescent lipids and 25 µM chloroquine. When incubating with fluorescent GSLs, the lipids can preferentially partition into serum proteins rather than cellular surfaces; either reduced-serum or serum-free medium have been used to enhance uptake. The cell culture medium used to grow the HCT 116 MCTSs was removed and replaced with either 200 µL of reduced-serum cell culture medium containing the fluorescent GSLs or 200 µL of reduced-serum cell culture medium containing the fluorescent GSLs and chloroquine. Thirty spheroids were treated with fluorescent GSLs for 24 hours. Half were treated with fluorescent GSLs alone and half with GSLs plus chloroquine. After incubation, the spheroids were removed from the 96-well plate and washed with serum-free medium to remove excess GSLs, chloroquine, and serum prior to serial trypsinization and single cell isolation.

2.3.3 Single Cell Isolation

Figure 2.1 presents an overview of the experiment. Fifteen spheroids treated only with the fluorescent GSL substrates were subjected to serial trypsinization [8]. The spheroids were incubated with chilled 0.05% trypsin/EDTA for 3 min at room temperature with gentle agitation to dislodge intact cellular layers. After 3 min, chilled media with serum was added to quench proteolysis. The wash medium was collected
containing a grouping of cellular layers. The serum-free medium washing, trypsinization, and serum-containing medium washing steps were repeated. Layers from the first three treatments were pooled and labeled as the outer surface of the spheroids (Figure 2.1-orange, referred to as outer). Cell layers from three subsequent treatments (fractions 4-6) were pooled representing an intermediate quiescent middle region of the spheroids (magenta, labeled as middle). The dead cells and necrotic cell debris (grey) were discarded, and the remaining cells, representing pre-necrotic cells (cyan, labeled as core) were selected as the third group. Each of these populations of intact cells was washed three times with PBS. Each group of cells was tritivated by rapidly pipetting the solution to produce a single cell suspension. Approximately 1/3 of the suspensions were each incubated with 4% paraformaldehyde in PBS to fix cells; the other 2/3 were homogenized and stored at -80 C°. Fixing was halted using a wash of 10 mM glycine in PBS followed by two additional washes of PBS. Single cells were then resuspended in PBS and stored at 4 C° until analysis. The procedure carried out in parallel for fifteen spheroids incubated with fluorescent GSLs and chloroquine.
Figure 2.1: Experimental overview. HCT 116 MCTSs were simultaneously incubated with GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665. After 24h, the spheroids underwent serial trypsinization, yielding sheet of cells from the outer (orange), middle (magenta), and core (cyan) regions. Dead cells and cellular debris are contained within the center (gray); this material was discarded. These cellular sheets underwent trituration to isolate individual cells, and the cells were then fixed with paraformaldehyde (PFA). After fixation, cells were analyzed with CE-LIF.
2.3.4 CE-LIF

The three-color single-cell CE system has been described elsewhere.22 Solid-state lasers produced excitation at 473, 532, and 633 nm. Detection used a sheath-flow cuvette [38]. The red, green, and blue laser beams were focused ~50 µm, 70 µm, and 90 µm downstream from the capillary terminus, respectively, to minimize photobleaching and spectral crosstalk [22, 29]. Fluorescence was detected in three spectral channels with single-photon counting avalanche photodiodes.

The separation buffer was 100 mM TRIS, 100 mM CHES, 20 mM sodium dodecylsulfate, and 5 mM α-cyclodextrin. 30-µm ID, 150-µm OD, 50-55 cm long uncoated fused silica capillary (Polymicro) were used for the separation. A Spellman CZE1000R high voltage power supply was used to drive the separation at 26 kV. A locally constructed Plexiglas injection assembly was used to introduce running buffer and single cells onto the capillary [27].

An Olympus IX70 inverted microscope was used for cell injection [27]. A 5 µL suspension of cells in PBS was placed onto a glass slide. The separation capillary, held in the injection assembly, was lowered with a micromanipulator until the capillary was a few micrometers above a cell of interest. A two-second pulse of negative pressure (11 kPa) injected a cell into the capillary. The capillary inlet was then placed in a solution of running buffer, and cells were incubated within the capillary for two minutes before the separation voltage was applied. A total of 212 cells were analyzed over a 30 day period after fixing. A new capillary was used at the beginning of each day and cells from the six conditions (outer, middle, and core, with/without chloroquine) were sampled randomly.
Cell diameters were measured using the inverted microscope equipped with a 2.0 megapixel camera, and accompanying software (Motic Group). The cells used for diameter measurements were from the same six subpopulations as those analyzed by CE-LIF, but were not the same cells that were electrophoresed.

A handful of cells were analyzed using 10 mW laser power, which sometimes resulted in a fluorescence signal that saturated the photodetectors. Out of the total 212 cells analyzed, data from 20 cells were discarded because of saturation, poor signal-to-noise ratios, poor peak resolution or shape, and traces that could not be properly aligned. A total of 192 single cell multicolor electropherograms were processed further.

2.3.5 Data Analysis

Data analysis was performed in MATLAB. Electropherograms were corrected for dead time of the avalanche photodiode treated with a five-point median filter to remove spikes, and smoothed by convolution with a 0.1 s wide Gaussian filter function [39-40]. A two-point algorithm was used to temporally align the electropherograms based on the migration of GM1-BODIPY-TMR and Cer-BODIPY-TMR.

BODIPY-FL generated a small amount of spectral crosstalk into the BODIPY-TMR channel (<1.5%); the BODIPY-TMR labeled GSLs exhibited less spectral crosstalk into the BODIPY-650/665 channel (~0.1%). Spectral crosstalk was removed by subtraction.

Peak identities were based on the known migration pattern of BODIPY-labeled GSLs [21-22]. Peak areas were estimated by fitting each peak with a Gaussian function. The overlapping GlcCer and Cer peaks were fit with the sum of two Gaussian functions. To correct for differences in cellular uptake and laser powers, each electropherogram
was normalized by the total fluorescence peak area in each spectral channel. All reported uncertainties are the standard error of the mean.

Principal component analysis was performed by first trimming each electropherogram to isolate the separation window (9 – 10.3 minutes). The three-color traces from each cell were then concatenated into a single vector. The BODIPY-TMR data was appended after the BODIPY-FL data followed by the BODIPY-650/665 data. The data were mean-centered by subtracting the average concatenated electropherogram from all electropherograms. Principal component analysis was performed using singular value decomposition [41].

2.4 Results and Discussion

2.4.1 Single Cell Isolation and Characterization

Diameters (~12.5 µm) were measured for a set of cells isolated from each region with and without chloroquine treatment (Figure A.2.1 in Appendix A). A One-Way ANOVA indicated that neither spatial location nor pharmacological treatment significantly altered cell size (p = 0.24). HCT 116 MCTSs have a radius of approximately 500 µm. The inner 250-µm radius is a fully necrotic core region consisting of dead cells and cellular debris while the outer 250 µm consists of concentric spherical cellular layers [4-5]. In one experiment, up to 10 rounds of serial trypsinization were performed before reaching the fully necrotic core, at which point the spheroid architecture disintegrated. Assuming equal effectiveness, each trypsinization iteration removed a ~25 µm layer, corresponding to a layer two cells thick. This result is consistent with the initial report of
serial trypsinization, which determined that dissociated layers of mouse mammary tumor MCTSs were 24-µm thick, and with subsequent reports on the technique [9-11].

The first three iterations of serial trypsinization isolated cells from the outer 75 µm (outer proliferative region) of the HCT 116 MCTSs, approximately six concentric layers thick. The middle region contained cells from the next three iterations of serial trypsinization, corresponding to six more concentric cellular layers at a depth of 75-150 µm. The pre-necrotic core region at a depth of 150-250 µm consisted of approximately eight concentric cellular layers. The dead cells and debris were discarded.

2.4.2 Three Color Electropherograms

A three-color electropherogram of the substrate-containing incubation medium is presented in Figure A.3.1 (Appendix A). Only trace impurities were noted. Separation efficiencies exceeded 200,000 theoretical plates. A total of 192 single cells generated signals within the linear range of our photodiodes: 33, 32, and 35 cells were analyzed from the outer, middle, and core regions of naïve HCT 116 MCTSs (respectively) and 29, 30, and 33 cells were analyzed from the outer, middle, and core regions of chloroquine-treated HCT 116 MCTSs.

Figure 2.2 presents the three-color electropherograms generated from a set of 29 cells isolated from the outer layer of the MCT 116 MCTS. Electropherograms generated from other portions of the MCTS and from cells treated with chloroquine are presented as Figures A.4.1 and A.4.2 (Appendix A). In general, components are well resolved, and the electrophoretic pattern is quite consistent for each substrate.
2.4.3 Effect of Chloroquine on Fluorescent GSL Uptake

Cells treated with chloroquine appeared to generate much larger signals than untreated cells. Assuming that metabolism does not change the spectral property of the dye, the total fluorescence signal within a spectral channel of an electropherogram is proportional to the amount of fluorescent substrate taken up by the cell. Figure A.5.1 (Appendix A) shows that treatment significantly (p<0.05 to p<0.001) increased uptake of the substrates in the outer and middle, but not the core regions of HCT 116 MCTSs.

This increase in uptake is consistent with chloroquine’s known short-term (under 2 hr) increase in uptake of exogenous species including lipids; at longer times, chloroquine can also inhibit uptake [32, 42]. Differences between fluorescent GSL and chloroquine transport presumably resulted in the observed spatial distribution of uptake of the fluorescent GSLs, where treatment significantly increased uptake only within the outer and middle regions of the HCT 116 MCTSs. The culture medium containing the fluorescent GSLs and chloroquine would be more accessible to the outer...
and middle regions of HCT MCTSs. If chloroquine transport to cells within the core region lags that of the fluorescent GSLs by 2 hr or more, chloroquine will not increase uptake in the core region, yet would still retain its action as a potent catabolic inhibitor.

2.4.4 Metabolic Cytometry of HCT 116 MCTSs

Electropherograms plotted as the mean ± one standard error of the mean are shown in Figure 2.3A, 2.3B, and 2.3C for the untreated cells; Figure 2.3D, 2.3E, and 2.3F expand the vertical axis to highlight low-level metabolites. Average electropherograms for the chloroquine-treated cells are shown in Figure 2.3G, 2.3H, and 2.3I; Figure 2.3J, 2.3K, and 2.3L present enlargements.

During the 24 hour incubation, fluorescent GSLs were taken up, transported, and metabolized in all regions of the HCT 116 MCTSs. Single cells within all regions, regardless of pharmacological treatment, showed both anabolism and catabolism of the GM3-BODIPY-FL and LacCer-BODIPY-650/665 substrates. GM1-BODIPY-TMR primarily exhibited catabolism. Occasionally, some anabolic products (likely GD1a or GD1b based on their migration time) generated peaks that migrated before GM1-BODIPY-TMR, but these peaks did not appear in the majority of electropherograms and when they did appear, their heights were near the noise floor [21].
Figure 2.3: Average multicolor metabolic cytometry electropherograms, normalized to unit area, for each region within the spheroids. The line width encompasses ± standard error of the mean for all cells in each location and pharmacological condition. A, B, and C compare GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) by location in HCT 116 MCTS single cells. D, E, and F present enlargements of the Y-axis of A, B, and C (respectively). G, H, and I compare by location in chloroquine (CQ)-treated MCTS. J, K, and L present enlargements of the Y-axis of G, H, and I (respectively). Numbers indicate the metabolic product (1-GM1, 2-GM2, 3-GM3, 4-LacCer, 5-GlcCer, and 6-Cer). Red numbers indicate the fluorescent substrate in each channel. An asterisk indicates a significant difference (p < 0.05) in peak area using a One-Way ANOVA as described in the text.
LacCer-BODIPY-FL, the first catabolic product of GM3-BODIPY-FL, was the most abundant BODIPY-FL labeled species observed in untreated cells, which indicates that GM3-BODIPY-FL underwent rapid catabolism. The GM1-BODIPY-TMR substrate and its catabolic product GM2-BODIPY-TMR dominated the BODIPY-TMR traces. The LacCer-BODIPY-650/665 substrate was the most abundant component within the BODIPY-650/665 traces. LacCer was the most abundant species in both the BODIPY-FL and BODIPY-650/665 channels. The accumulation of LacCer suggests that its metabolism was slower than that of the other GSL species in these HCT 116 cells.

Peak areas for each analyte were calculated and a one-way ANOVA was used to determine whether cellular localization within naïve HCT 116 MCTSs influenced the presence of specific GSL species. Asterisks located above peaks in Figure 2.3A, 2.3B, and 2.3C indicate a significant difference (p<0.05) in GSL metabolism between the three regions. All GM3-BODIPY-FL metabolites showed differential abundance as a function of location, while substrate abundance did not significantly change. The substrate GM1-BODIPY-TMR as well as its metabolites GM2-, LacCer- and GlcCer-BODIPY-TMR showed differential abundance within the MCTS architecture. Finally, the substrate LacCer-BODIPY-650/665 and its metabolites GM1-, GM2-, GlcCer- and Cer-BODIPY-650/665 exhibited differential abundance with location.

The procedure was repeated for chloroquine-treated HCT 116 MCTSs. Again, all GM3-BODIPY-FL metabolites demonstrated differential abundance while the GM3-BODIPY-FL substrate was unchanged. GM1-, GM2-, LacCer-, GlcCer-, and Cer-BODIPY-TMR showed differential abundance with cellular location, while GM3-BODIPY-TMR did

The traces in Figure 2.3 were recompiled to better illustrate the effect of chloroquine on GSL metabolism as a function of spatial location, Figure 2.4. Student’s t-Tests on peak areas were used to determine whether it significantly altered GSL metabolism. With the exception of GlcCer-BODIPY-650/665 within the middle region, the drug molecule significantly altered the production of all catabolic products from all three fluorescent substrates in single cells isolated from all three spatial locations within HCT 116 MCTs (p < 0.05). Nearly all of those differences were reductions in GSL abundance with the exception of LacCer-BODIPY-FL, GM2-BODIPY-TMR, and LacCer-BODIPY-650/665, which increased in abundance after treatment.
Figure 2.4: Effect of chloroquine treatment on fluorescent GSL metabolism by region. The data in Figure 3 are recompiled to illustrate the effect of chloroquine (CQ) in different regions. Each colored trace is shaded to encompass + standard error of the mean for all cells in a specified location and pharmacological condition. A, B, and C show the effect of CQ on GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells from HCT 116 MCTSs located within the outer region. D, E, and F show the effect of CQ on GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) within the middle region. G, H, and I show the effect of CQ on GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) within the core region. Numerical indices are the same as those depicted in Figure 3. An asterisk indicates a significant difference (p < 0.05) in peak area using a Student's t-test as described in the text.
Catabolism of GSLs mainly occurs within the lysosome, yet the lysosomal inhibitor chloroquine caused an increase in the relative abundances of a few catabolic metabolites. There are three possible reasons for this effect. First, it is possible that some GSL catabolic enzymes were more pH sensitive than others, since chloroquine dramatically increases intralysosomal pH [31]. Second, the drug could have targeted selected enzymes responsible for GSL catabolism, much as it targets selected lysosomal hydrolases [32]. The third possibility is kinetic, which would imply rapid catabolism of GM3-BODIPY-FL and GM1-BODIPY-TMR. If the transport and catabolism of these substrates were more rapid than the transport and pharmacological action of chloroquine, a buildup of some catabolic products in cells would be expected as the kinetics of GSL catabolism outran the kinetics of chloroquine action. When the drug molecule does take effect, subsequent catabolism would be quenched.

The observed increase in the relative abundance of LacCer-BODIPY-650/665 upon treatment again suggests that LacCer catabolism was relatively slow. GSLs initially reside in the cell membrane and rely on recycling to enter cellular compartments [43-44]. Therefore, the increase in relative abundance of LacCer-BODIPY-650/665 may be related to a decrease in intracellular trafficking of cell surface lipids produced by chloroquine [33]. Since GM3-BODIPY-FL and GM1-BODIPY-TMR were not affected in the same way, this hypothesis is unlikely unless the effects of chloroquine involve selective targeting of LacCer-BODIPY-650/665 enriched portions of the cellular membrane.

The effect of chloroquine on the relative abundance of each fluorescent GSL species is summarized in Figures A.6.1, A.6.2, and A.6.3 (Appendix A). This view of GSL
metabolic abundance reiterates the information shown in Figure 2.3 and Figure 2.4, highlighting the variation in GSL metabolism as a function of both cellular location and pharmacological treatment within single cells isolated from HCT 116 MCTSs.

2.4.5 Quantifying Cell-to-Cell Variation in GSL Metabolism

Each cell produced a three-color electropherogram that contains information on that cell’s GSL metabolism, starting from three different points in the metabolic cascade. We used principal component analysis to characterize the cell-to-cell heterogeneity in that metabolism. We first reduced dimensionality by concatenating the three-color data array into a one-dimensional vector for each electropherogram as described in the experimental section.

Principal component analysis was performed on the concatenated electropherograms. Six principal components were retained as determined by a Scree plot [45], corresponding to 91.5% of the total information content within the 192 electropherograms. A score plot of the first and second principal components is given in Figure 2.5A with the single cell electropherograms grouped by condition. Each point represents an entire three-color electropherogram taken from a single cell. The scatter shown in Figure 2.5A highlights the rich cellular diversity present in GSL metabolism.
Figure 2.5: Principal component analysis of multicolor single cell electropherograms. A. PC1 versus PC2 score plot of the concatenated multicolor electropherograms for each cell in each condition. B. Plot of the Euclidean Norm of PCA scores between each point within a specified cluster and the multivariate spatial median for all points within the specified cluster. Error bars are ± one standard error of the mean.

Cells from the outer, middle, and core regions of untreated HCT 116 MCTSs tended to overlap in the score plot. Cells from the outer (•) and middle (x) regions of chloroquine-treated MCTSs overlapped but were distinct from cells from the other conditions. This result indicated that, on average, treatment caused the GSL metabolism of single cells within the outer and middle regions of the HCT 116 MCTSs to more closely resemble one another and differ from cells within the core region. This result is similar to a report of spheroids treated with irinotecan.4 At 24 hours of irinotecan incubation, IMS revealed that the outer and middle regions of HCT116 spheroids showed similar mass spectral characteristics and the outer region gave a metabolically distinct small molecule profile compared to that of the central core.
In scatterplots such as Figure 2.5A, the data’s heterogeneity can be measured as the degree of scatter in the distance between points and the center of the cluster. The Euclidean norm (EN) is traditionally used to calculate distance between points.

Anderson proposed a distance-based test for population homogeneity of multivariate dispersions within multivariate principal component analysis space [46]. We used this distance-based test to determine if the degree of cell-to-cell variation present in GSL metabolism was affected by location or chloroquine treatment. First, principal component analysis was performed on the concatenated single cell electropherograms. Next, the scores were grouped by condition (outer, middle, and core, with and without chloroquine). This grouping yielded a j-by-k matrix S where j represented the number of single cell electropherograms within a condition and k represented the number of relevant principal components. ENs were then calculated between the scores from cells in each group and the center that group

$$\sum_{j=1}^{6} \left(S_{jk} - M_{jk}\right)^2$$

(Eq. 2.1)

where $M_{jk}$ represented the median score for all j samples along a specific principal component k. Multivariate means (centroids), multivariate trimmed means, and multivariate spatial medians can all be used to estimate the multivariate center of a data cluster, but the multivariate spatial median is a more robust measure that makes no assumption of the underlying statistical distribution of the data [47].

This analysis yielded six groups of distances (one for each condition investigated) indicating the degree of heterogeneity in single cell GSL metabolism. A bar graph of
these distances is shown in Figure 2.5B. A One-Way ANOVA was then used to test whether spatial location or treatment significantly altered the spread/heterogeneity of single cell electropherograms. A p value of 0.45 was obtained, indicating that the cell-to-cell variation of GSL metabolism did not change with location within the HCT 116 MCTS or with chloroquine treatment.

This result does not indicate that GSL metabolism does not vary within single cells; the single cell traces shown in Figures A.5.1 and A.6.1 (Appendix A) as well as the score plot shown in Figure 2.5A clearly illustrate metabolic heterogeneity. Instead, this result indicates that neither spatial location nor treatment significantly alter the degree to which single cells exhibit GSL metabolic heterogeneity.

2.5 Conclusions

We report several developments in single cell metabolic analysis. This is the first report of CE with three-color LIF detection for single cell analyses. This is the first report of the use of metabolic cytometry to analyze single cells within a biological system that mimics the development of microenvironments similar to those found in tumors. Metabolic cytometry has been used to monitor GSL metabolism in cells isolated from primary neuronal tissue [27-29], but the tissue was triturated into single cells before GSL incubation, losing all information on the cellular position within the complex tissue. Maintaining the tissue’s architecture during incubation allowed us to probe the inherent differences in the GSL metabolome as a function of location on a single cell level. We also successfully analyzed pharmacological effects of the lysosomal inhibitor and
potential therapeutic chloroquine on a single cell level, which consistently resulted in decreased catabolism for all substrates and spheroid regions.

Finally, coupling principal component analysis with a distance-based heterogeneity measure allows is a simple, elegant, and powerful way to statistically quantify and compare cell-to-cell variability on a multi-faceted whole-spectrum level. Traditional measures of single cell heterogeneity in flow cytometry rely on coefficients of variation from the measurement of a few discrete parameters [48]. Here, we characterize heterogeneity based on three-color electropherograms, dramatically increasing the information content of the measurement. To our knowledge, this is the first use of such an approach to characterize cell-to-cell variation and could be readily expanded to include other measurements, including optical and mass spectrometry, surface analyses, etc.

2.6 Acknowledgments

The authors acknowledge Dr. Ole Hindgaul and Dr. Monica Palcic for their gift of the fluorescent GSL substrates and Dr. Ronald Schnaar and Dr. Jill Prendergast for helpful discussions. Funding was provided by the 2011 Starter Grant from the Society for Analytical Chemists of Pittsburgh and The University of Notre Dame (A.B.H.) and grant R01NS061767 from the National Institutes of Health (N.J.D)

2.7 References


CHAPTER 3: 
HIGH RESOLUTION ANALYSIS OF THE LIPIDOME OF THREE DIMENSIONAL CELL CULTURES

This chapter describes an ongoing collaboration between Dr. Hummon’s lab and Dr. Theodore Alexandrov’s lab. The initial studies were performed at Bruker Daltonics life science imaging headquarters in Bremen, Germany in the spring of 2013 with the assistance of Dr. Michael Becker and Dr. Jens Fuchser. Dr. Andrew Palmer, a postdoctoral researcher in Dr. Alexandrov’s lab, performed the database searches and rendered the ion images collected in Bremen into multiple platforms to begin building a web-searchable database of metabolites and lipids found in the spheroids. Another round of experiments was performed in the summer of August 2015 at Bruker Daltonics world headquarters in Billerica, MA. Dr. Katherine Kellersberger assisted with sample preparation, data collection and some data analysis. In this study, I was responsible for growing and maintaining the spheroids, sectioning and imaging mass spectrometry sample preparation. I also assisted Dr. Palmer in interpreting metabolic and spatially resolved data gathered for the online database.
3.1 Abstract

Here, we present spatially resolved direct injection electrospray ionization (ESI) MS/MS sequencing as an orthogonal technique to MALDI IMS of lipids distributed in drug-treated three-dimensional cell culture models of the colorectal adenocarcinoma cell line HCT 116. Control and three-dimensional cell cultures treated with the drug chloroquine were analyzed. Isotopic peak patterns were identified in high resolution IMS images of HCT 116 spheroid serial sections. Putative molecular identifications based on these isotopic patterns were used to examine lipid species within distinct, biologically relevant regions of the cell culture models. The hypotheses generated from the IMS data were evaluated with MS/MS sequencing of serially trypsinized fractions of the cell culture models, which correspond to biologically relevant spatial regions within in these tumor models. Together, these combined methods represent a cohesive progression from a semi-targeted, discovery based analysis to a thorough examination of changes in lipid metabolism across different microenvironments of complex tumor mimics.

3.2 Introduction

Since its inception in the late 1990s, imaging mass spectrometry (IMS) has traditionally been applied to molecular imaging of tissues and tumor biopsies from both human clinical samples and lab animals [1-3]. In recent years, there has been a push to apply IMS to analyze a diverse array of samples, especially cell culture models such as
two-dimensional co-culture models and more complex three-dimensional models such as the spheroids described in the present study [4-8].

In the experiments presented here, we use high spatial and very high mass resolution Fourier transform ion cyclotron resonance mass spectrometry (FT-ICR MS) to image sections of three dimensional cell culture (3DCC) models of the HCT 116 colon carcinoma cell line. The mass measurements were performed with seven and twelve Tesla Bruker SolariX FT-ICR mass spectrometers equipped with MALDI sources. With the high mass resolving power (<100,000 at 700 \textit{m/z}) and 20 µm spatial resolution in the MALDI source, an average of 2,500 spectra/sample was acquired in the 100-1500 \textit{m/z} range. This study represents the highest mass resolution IMS data acquired from 3DCC models to date. This allowed us to examine lipid and small molecule distributions both in untreated spheroid sections as well as in sections of spheroids after six, twelve and twenty-four hour treatments with the small molecule chloroquine.

Lipids represent an extraordinarily important class of biomolecules that have numerous critical functions within cells [8, 9]. In addition to being the principal component of cell membranes, the structural complexity of lipid molecules allows them to be involved in a diverse array of cell signaling functions. The ability to store potential energy in long chains of hydrocarbons makes lipids important in cellular metabolism and energy storage. Different cell types within the body have different lipid profiles, which are altered in many types of diseases, including cancer [9].

Aberrant metabolism as first described by Otto Warburg is a hallmark of cancer progression and has been studied in depth in recent years [10, 11]. Deregulated
metabolism allows cancer cells to survive environments that would otherwise cause normal cells to die, such as hypoxic zones or areas scarce in nutrients [12]. The energy storage function of lipid molecules becomes critical in populations of cells exposed to metabolic stress that may inhibit other energy producing pathways in the cell such as glycolysis or the citric acid cycle. These cells are known to undergo mitochondrial de-coupling and to switch to aerobic glycolysis which provides both ATP and, more importantly, carbon skeleton intermediates for the increased production of nucleic acids, lipids and other cellular building blocks [13-15]. Autophagy is a cellular process used in non-malignant cells as a way of recycling long-lived proteins and damaged organelles through the lysosome that may work in conjunction with the ubiquitin-proteasome system [16]. Autophagy is also likely one of the principal ways cancer cells are able to cope with the metabolic stress of a quickly growing tumor microenvironment [16, 17].

As described previously, the three different populations of cells existing within fully developed 3DCC spheroids are in different metabolic and proliferative states [4, 7, 18]. For this study, spheroids were treated for varying amounts of time with chloroquine, an anti-malarial drug that inhibits lysosome-mediated cellular catabolism pathways, presumably by changing the pH of lysosomes and preventing acidic degradation of lysosomal contents [19]. We hypothesize that chloroquine treatment will change cellular metabolism throughout the entire spheroid structure, but its impact will be the greatest in the hypoxic necrotic core of the spheroids where cells must rely on autophagy and other alternative metabolic pathways for survival [20]. By inhibiting the
autophagic processes, the innermost cells of the spheroid will be forced to change their metabolic profile in favor of processes such as β-oxidation, potentially rendering them more sensitive to conventional chemotherapeutic approaches, or resulting in increased cell death independent of additional treatment. The ability to target and destroy different populations of cancer cells will lead to more effective cancer treatment and could prevent relapse in the future by eliminating cancer cell populations that normally evade destruction in conventional treatments [19, 21].

3.3 Materials and Methods

3.3.1 Cell Culture

The HCT 116 colorectal carcinoma cell line was obtained from ATCC (Manassas VA, USA) and maintained at 37°C and 5% CO₂ in McCoy’s 5A media (Gibco, Gaithersburg MD, USA) supplemented with 10% fetal bovine serum (Hyclone, Logan UT, USA) and 5% L-Glutamine (Gibco). These cell lines were used within three months of resuscitation of frozen aliquots thawed from liquid nitrogen. The provider assured the authentication of these cell lines by cytogenetic analysis. To induce three-dimensional growth, the inner 60 wells of a 96 well plate (Thermo Scientific, Carlsbad CA, USA) were coated with 0.2% agarose (Sigma Aldrich, St. Louis MO, USA) dissolved in the cell culture media. Six thousand cells/well were seeded into these agarose-coated wells and allowed to grow for two weeks, with full media changes at 4, 8 and 10 days and the day before drug treatment. For chloroquine (Sigma) treatment, the drug was first dissolved in purified
water and then diluted into cell culture media at 25 µM. Cells were treated for 6, 12 and 24 hours.

3.3.2 Sectioning and MALDI Sample Preparation

Spheroids were removed from the growth plates after treatment and embedded in gelatin (Kraft Foods, Chicago IL, USA) for cryosectioning as previously described [4]. Briefly, gelatin was heated to dissolve in purified water and a base layer of gelatin was applied to the bottom of a 12 well cell culture plate (Thermo) and allowed to cool to room temperature. The spheroids were removed from the growth plate at the end of the treatment period and rinsed in 1x PBS (Gibco). Rinsed spheroids were then placed on the solidified gelatin and covered with a layer of warm gelatin and immediately placed at -80°C until sectioning.

For sectioning, the gelatin blocks containing the embedded spheroids were scooped from the cell culture with a metal spatula and placed onto sectioning mounts in water and allowed to equilibrate to -30°C in the cryomicrotome (Leica Biosystems, Nussloch, Germany). Spheroids were sectioned into 14 and 16 µm thick sections and thaw-mounted onto room temperature ITO-coated glass MALDI slides (Delta Technologies, Golden CO, USA). Once mounted, the slides were immediately transferred to a vacuum desiccator and stored until matrix application.

MALDI matrix was applied by sublimation using an in-house built sublimation chamber for experiments on the 12T instrument while a TM Sprayer (HTX Technologies, Chapel Hill NC, USA) was used for the 7T experiments. The matrix used was 2,5-dihydroxybenzoic acid (DHB) (Bruker Daltonics, Billerica MA, USA). Matrix thickness was
determined by an optical sensor mounted in the sublimation apparatus.

Dihydroxybenzoic acid was dissolved in 90:10 Ethanol:H₂O (Burdick and Jackson, Phoenix, AZ) mixture and set to run at 200 µl/min into the injection loop of the TM Sprayer. The nozzle was heated to 60°C and set to make 8 passes across the imaging slides with every other pass offset by 1.5 mm. Resulting matrix coats were homogeneous and reproducible for each slide.

3.3.3 MALDI Imaging

MALDI spectra were acquired from the prepared slides using a 7 or 12T SolariX XR MALDI-FT-ICR (Bruker Daltonics, Billerica MA, USA) using FlexImaging 4.1 (Bruker) image acquisition software. Laser raster width was set to 20 µm and laser power was adjusted manually prior to each image acquisition. Spectra were acquired from 100-1500 m/z in positive ion mode.

3.3.4 Serial Trypsinization

After the appropriate amount of treatment time either with or without chloroquine, spheroids were removed via serological pipet from their growth plate (n=60 spheroids/plate) and rinsed thoroughly in 37°C 1x PBS (Gibco). PBS was aspirated and discarded and the serial trypsinization procedure was carried out as described in McMahon et al. [22] Briefly, 2.5 mL of 0.05% trypsin was added to the spheroids which were then agitated on a rocking platform for 3 minutes. After agitation, cold complete McCoy’s 5A media (Invitrogen) with 10% FBS (Gibco) was added and the spheroids were agitated for an additional minute. After 1 minute with the media added, the agitation
was stopped and the supernatant was collected without disturbing the spheroids. The spheroids were then rinsed and agitated for 1 minute with warm media not containing FBS. The supernatant was collected and saved. This entire process was repeated between 6 and 8 times for each plate of spheroids. Three or four treatments were pooled and counted as one region of spheroids (i.e. the first three treatments were pooled as the “Outer” fraction, the next three treatments were pooled as the “Intermediate” fraction and the remains after six washes were kept and pooled as the “Core” fraction). Pooled fractions were centrifuged at 4°C and washed three times with HPLC grade H₂O (Burdick and Jackson), flash frozen in liquid nitrogen and stored until lipid extraction was performed.

3.3.5 Lipid Extraction

Cell fractions from serial trypsinization were brought to room temperature slowly on ice and briefly centrifuged. Cells were resuspended in HPLC grade methanol (Burdick and Jackson) and transferred to 2 mL glass vials for a modified Folch lipid extraction based on Ekroos et al.[23] Lipid standards PC (17:0/17:0), PE (17:0/17:0), PS (17:0/17:0) and LysoSM (d17:1) (Avanti Polar Lipids, Alabaster AL, USA) were spiked into each sample after the addition of the methanol for a final concentration of 2 pmol/µl. Post-extraction, lipid extracts were dried in a vacuum concentrator (Eppendorf, Mannheim, Germany) and stored at -20°C until mass spectrometric analysis.
3.3.6 ESI-MS/MS Sequencing of Lipid Extracts

Lipid extracts were resuspended in a 1:2 mixture of methylene chloride:methanol with 5 mM ammonium chloride (Sigma) in a final volume of 350 µl. Each fraction of spheroids was analyzed in technical triplicate. Samples were run for 3 min injections using a syringe pump (Thermo) set to 10 µl/min coupled directly to a Q-Exactive mass spectrometer (Thermo). Initial analysis was done in negative ion mode using a Top 12 data-dependent MS/MS method with dynamic exclusion set to 40 s and using stepped normalized collision energy centered at 30 with steps of 33.3%.

3.3.7 Peaking Picking and Database Searching of MALDI Data

MALDI spectra were extracted from their native format and subjected to the following workflow:

![Figure 3.1: Conceptual overview of “Spatial Metabolomics Pipeline” as applied to complex high resolution IMS datasets. Adapted from Dr. Andrew Palmer.](image)

This portion of the workflow was performed at the University of Bremen in Germany by Dr. Andrew Palmer. The results from the peak picking and HMDB searches were compiled into a searchable database using the Java (Sun Microsystems, Santa Clara CA, USA) coding language. This protocol allowed for easy comparisons among and
across time points for each search hit generated from a peak selected in the MALDI data sets.

3.3.8 Database Searching of ESI-MS/MS Lipid Data

Data files in the “.raw” format obtained from the mass spectrometer were first manually examined for the presence or absence of lipid standard peaks in the MS spectra and then examined for headgroup peaks and other fragmentation data in the MS/MS spectra. Post-manual analysis the “.raw” files were converted to “.mzXML” files using ProteoWizard 3.0 (freely available at http://proteowizard.sourceforge.net/). Peak lists were further extracted from the mzXML files and searched against either the LIPID MAPS Lipidomic Gateway (http://www.lipidmaps.org/) or the Human Metabolome Database (HMDB) using scripts written by Dr. Palmer [24, 25]. Search results were compiled into an Excel spreadsheet and results were manually compared to the raw data to check for correct matches.

3.4 Results and Discussion

3.4.1 Spheroid Metabolic Analysis

The most abundant analytes found in the 100-1500 Da mass range are lipids. As mentioned in the introduction, lipids are extremely important molecules within the cell [8]. Outside of making up the cell and various organelle membranes, lipid molecules also have a variety of functions including energy transport, storage and cell signaling. Many normal lipid pathways can be altered in the tumor environment, including lipid
signaling, transport and metabolism [9, 13]. A particularly dominant trend examined in this experiment is changes in lipid metabolism in each region of the spheroids over a time course with drug treatment. Cells with access to fresh nutrients typically rely on canonical energy producing pathways such as glycolysis, the TCA cycle and mitochondrial oxidative phosphorylation [11]. As enumerated in the introduction, cancer cells are able to subvert these normal processes in favor of aerobic glycolysis and a higher production of carbon skeleton intermediates. However, when glucose becomes scarce or unavailable as in poorly vascularized regions of tumors, mimicked here by the necrotic core region of the tumor spheroids, cells are forced to rely on alternative forms of metabolism for survival such as β-oxidation of lipid molecules or autophagy [15, 26]. As depicted in Figure 3.2, saturated lipid species are more abundant in the outer layers of untreated spheroids, consistent with these cells having access to fresh nutrients in the cell media during growth and therefore likely relying on aerobic glycolysis for their primary source of energy and carbon intermediates for the synthesis of other biomolecules [15, 17].
Figure 3.2: Metabolic trends from different spheroids at different time points of chloroquine treatment. 706.5381 m/z corresponds to a completely unsaturated lipid species. The data indicates that this particular lipid species is most likely not being used as an energy source in the outer population of cells at each time point. Conversely, 766.5745 m/z indicates a partially saturated lipid species. The localization of this species implies that the inner populations of cells in the spheroid at each time point are being forced to rely on β-oxidation as their primary energy source due to lack of other metabolites. There is a considerable increase in the amount of saturated lipid species in the core of 24 hour spheroid.
Lipid Metabolism in Spheroids

Untreated Spheroid (504 μm)  12 hour Treated Spheroid (504 μm)  24 hour Treated Spheroid (384 μm)

Brightfield

706.5381 m/z  
HMDB07869  
PC(14:0/16:0)

766.5745 m/z  
HMDB08914  
PC(18:3(6Z,9Z,12Z)/P-18:1(11Z))

Merge:  
706.5381 m/z-Green  
766.5745 m/z-Red
Examining the metabolic trends in the spheroid core, there is a noticeable upregulation of increasingly unsaturated lipid species, indicating that as cells in the tumor mimics become further removed from fresh nutrients, they are increasingly forced to rely on alternative metabolic pathways for energy in their hypoxic environment [26]. Finally, there is a very strong co-localization of semi-unsaturated lipid species with several different forms of carnitine which is heavily involved in transporting lipids across the mitochondrial membrane for use in β-oxidation as an alternative carbon source to sugars (Figure 3.3) [26, 27]. This trend is most prominent in the intermediate region of the spheroids, an area of the tumor mimics which is not yet fully hypoxic but has limited access to the cell media. These cells are still mostly alive with intact membranes and organelles, and rather than actively proliferating, are in a quiescent state [18]. The absence of strong carnitine accumulation in the core region of the spheroids is hypothesized to be a consequence of many of the cells in those regions being apoptotic or necrotic.
Figure 3.3: Carnitine Distribution in two different untreated spheroids at similar longitudinal depth in each spheroid. Carnitine is shown to be distinctly located in the intermediate region of these spheroids. This data could signify an upregulation of β-oxidation pathways in these regions. The multiple types of carnitine present likely indicate different lipid species being involved in these pathways.

3.4.2 Chloroquine Treatment

The goal of this study was to analyze how the anti-malarial and lysomotropic drug molecule chloroquine might affect the cells in each region of the spheroid. Figure 3.4 shows the drug distribution detected in different spheroids at each time point examined in this study. Also mapped is the distribution of the diprotonated form of the chloroquine molecule which follows a similar distribution to the monoisotopic parent molecule.
Chloroquine inhibits lysosomal degradation by entering the lysosome and accepting a proton which subsequently changes the pH of the lysosomal compartment. This not only inhibits specific enzymes within the lysosome but also prevents the drug molecule itself from being pumped back to the cytoplasm, thus causing a disruption of lysosome-mediated degradation pathways [28]. The ability to detect the protonated form of chloroquine is essential to this study as that is believed to be the “active” form of the molecule. As demonstrated in Figure 3.4, the molecule is found first gradually through the regions of the spheroid directly exposed to the cell culture media at the
twelve hour time point while the molecule is found nearly everywhere throughout the spheroid after 24 hours of treatment. Both molecules appear to be at a higher concentration in the outer regions of cells owing to the fact that those populations are composed of living cells which would still be intact and have normally functioning organelles. Since a smaller portion of cells in the necrotic core are believed to be intact there is less accumulation of the chloroquine molecule in that region.

3.4.3 Establishment of a Metabolomic Pipeline

IMS experiments can produce extremely complex datasets that contain large amounts of information [6, 8]. Software packages that accompany imaging instruments provide tools to view and analyze the molecular distributions captured in an image that work well for checking the localization of a few known peaks to assure reproducibility and quality from image to image. An individual researcher attempting to catalogue and observe the complex interactions of the hundreds or thousands of analytes in a single dataset, much less across multiple experiments, would take an inordinate amount of time using this software [7]. This situation is further complicated by the fact that MALDI IMS does not currently provide reliable identification methods for the analytes observed. Through the expertise provided by Dr. Theodore Alexandrov’s group, we began a collaboration to design and implement a “Spatial Metabolomics Pipeline” that would address the aforementioned problems regarding IMS dataset analysis (Figure 3.5). This pipeline accomplishes four main goals: 1) Provide putative identification of the analytes detected in high mass resolution (<100,000 k resolving power) IMS experiments based on the isotopic peak distributions present in the spectra, 2) Combine this
identification information with easy-to-use visualization software that retains the intrinsic spatial advantage of IMS experiments, 3) Create a tool that allows a researcher to easily visualize changes in analyte localization across multiple experiments (i.e. changes associated with a time course drug treatment), and 4) Cross-reference the putative database search hits with a complimentary identification method, in this case MS/MS sequencing performed on an Orbitrap mass spectrometer.

![Database Driven Spatial Metabolomics](image)

**Figure 3.5: Workflow overview of “Spatial Metabolomic Pipeline.”** The pipeline combines IMS data, database searches and conventional MS/MS sequencing data to assist in the analysis of highly complex datasets. Adapted from work done by Dr. Andrew Palmer.

Rendering the images with JavaScript allows for a different visualization scheme. The dimensionality of the data is also reduced to simple XML files so viewing the distribution of either multiple analytes across one image or a single analyte across multiple images is easily accomplished on almost any computer with a web browser.

Shown in Figure 3.6 is an example of two HMDB search hits across three time points of
spheroids treated with chloroquine. The top row is the [M+H]^+ peak, with the following rows being Sodium ("Na", [M+22]^+) and Potassium ("K", [M+44]^+) adduct peaks. Each column denotes a different region of the spheroid in each time point. The heat map indicates the abundance of a particular species in each region based on the intensities measured in the original IMS experiment.

Figure 3.6: The top row of boxes shows the JavaScript rendering of the distribution of HMDB08915. Represented are the Sodium adduct of PC(18:3(6Z,9Z,12Z)/P-18:1(9Z)) that is localized heavily to the core of 0 hr (Untreated) and 12h treated spheroids. The bottom row of boxes shows the distribution of the molecular ion corresponding to HMDB07876 (PC(14:0/18:3(9Z,12Z,15Z))) which shows a variable distribution across time points.
Transforming the IMS results into the JavaScript also allows different types of comparisons to be made, such as that in Figure 3.7, which shows the fold change abundance difference for each time point relative to the untreated sample.

**Figure 3.7: Log fold change compared to the Untreated samples of HMDB08915 demonstrating the magnitude of abundance differences between the treated time points.**

The use of JavaScript also allows the user to easily change the display settings such as the color scheme used for the heat maps, the angle of rotation (if rendering in 3D) and the scale on which the data will be displayed can be toggled between linear and logarithmic scales to help determine the magnitude of localization differences between the various samples.
3.4.4 Using ESI-MS/MS Sequencing Data to Compliment IMS Results

While each of the two mass spectrometric methods used here are quite useful on their own, when combined they offer unparalleled analytical power. For example, the spatially resolved nature of IMS is a valuable step in the analysis of any sample that contains biologically distinct regions; however, as it currently stands there is no reliable way to directly identify an unknown analyte \textit{in situ} during an IMS experiment. Various ESI techniques offer multiple fragmentation pathways and high resolution measurement of fragments that, coupled with well-curated fragment databases, result in high confidence identification of analytes. However, ESI sample preparation necessitates homogenization, digestion, extraction and destruction of the sample \cite{29-32}. The Spatial Metabolomics Pipeline aims to combine both of these powerful methods to offer a more comprehensive analysis of samples.

The MS/MS sequencing data was searched using the LIPID MAPS Lipidomic Database (www.lipidmaps.org) by extracting the fragment peak lists from the native data analysis program into .txt files and submitting them, either individually or through a batch process, to the LIPID MAPS database of glycerophospholipid precursor/product ions (- ion mode). Figure 3.8 shows the MS level summed mass spectra from each population of spheroid cells along with MS/MS fragmentation spectra and the structures of the lipids they are derived from using the LIPID MAPS Structure Database (LMSD) \cite{33}.

Spectra were considered a match for a particular lipid species only if they contained both parent ion and a second distinctive fragment corresponding to the loss
of a head group moiety from the parent ion or if the actual head group ion itself was detected. Finalized identification criteria have not yet been determined but previous lipid fragmentation studies are providing clues as to how to match fragmentation spectra to the correct lipid [9, 23, 29, 30].

Figure 3.8: Summed MS level spectra from one replicate of the direct injection ESI-MS/MS analysis of each population of cells from untreated spheroids. Insets are MS/MS spectra and the structures of the lipids they represent as well as the structural analysis obtained by submitting the search results to the LIPID MAPS online database.

While mass spectrometric studies of cell culture models are routine, in order to offer a viable comparison to the IMS data already acquired, the biologically important spatial distributions in the spheroid models need to be preserved. The serial
Trypsinization technique described in the methods section allows for different populations of cells to be separated from spheroids using dilute trypsin washes to be pooled and analyzed (Figure 3.9). A previous study by the Dovichi and Hummon groups showed that a three minute exposure to 0.05% trypsin removes two layers of cells corresponding to roughly 26 µm (assuming an average cell radius of 12.5 µm) [34].

By collecting and pooling the first three trypsin washes the “outer” population of cells was isolated. The cells produced from the subsequent three trypsin washes were collected and pooled as the “intermediate” region. Finally, the remaining cells were pooled as the “core” region. Each region was analyzed separately which allows for an accurate mapping of the lipids identified from the ESI experiment to each region of the spheroids thus making a direct comparison between the IMS data and ESI-MS/MS data possible.

Because ESI techniques do not inherently preserve the spatial distribution of analytes from samples, the data analysis presented another challenge in terms of
visually representing the results in their correct spatial distributions. Figure 3.10 represents one of the solutions to this problem produced by Dr. Palmer. Again, using web browser-friendly Java language and scripts these results are easily visualized outside of the laboratory on any standard PC without the aid of sophisticated data analysis software.

Figure 3.10: Visualization of direct injection ESI data shown in Figure 3.9 depicting the spatial distribution as well as the relative intensity in each region of three lipid species based on normalized intensities of each lipid parent ion detected. Peak intensities were normalized and averaged across three replicates collected from untreated spheroids. Adapted from work done by Dr. Andrew Palmer.
3.5 Conclusion and Future Directions

The goal of this study was to investigate chloroquine’s effects on each population of cancer cells in an in vitro model of advanced colon tumors. The effects of the drug molecule should be most pronounced in the core region of the tumor mimics as those cells are nutrient starved and rely on alternative metabolic pathways, including autophagy, to survive. We examined the distribution of lipids and other small molecules in 20 individual spheroids via high mass resolution IMS. To complement the IMS results direct injection ESI-MS/MS sequencing was used as an orthogonal technique to confirm the identification of various lipid species found in each biologically relevant region of untreated and treated spheroids [23].

This work has established a “metabolomics pipeline” that allows for the visualization of customized, searchable databases based on actual features found in individual IMS datasets. The coding and rendering features in this pipeline make it feasible to easily compare molecular feature distributions between different IMS datasets without the use of sophisticated data analysis programs or powerful data processing computers.

Chloroquine is currently involved in seven phase I/II clinical trials that are actively recruiting patients (www.clinicaltrials.gov), some examining it as an adjuvant therapy in late stage (III/IV) cancers of various types. None of the currently open studies involve the effects of chloroquine on colorectal cancer. Preclinical studies are needed in order to better establish the role the chloroquine actually plays in advanced tumors [19]. Studies such as this one seek to elucidate the mechanisms by which different
populations of cancer cells respond to autophagy inhibition. Autophagy is believed to be an especially important process by which cancer cells in hypoxic regions removed from nutrients are able to survive and evade some chemotherapy treatments.

Taken together, this collaboration has produced a tool that makes the analysis of hundreds of analytes across multiple time points possible without the need for complex algorithms, powerful data processing computers not normally accessible outside of a laboratory or a priori knowledge of every analyte present in the sample. While figures presented here were rudimentary, future iterations will expand the pipeline to server-based platform with advanced search and comparison features. Furthermore, the more advanced database will also compile data submitted from multiple research groups. Future versions will also have, where possible, spatially resolved MS/MS sequencing data accompanying the IMS data allowing for the identification of select analytes in the IMS images.

3.6 References


CHAPTER 4:
CHEMOMETRIC ANALYSIS OF MALDI MASS SPECTROMETRIC IMAGES OF THREE
DIMENSIONAL CELL CULTURE SYSTEMS

This chapter describes a workflow for the chemometric analysis of IMS images. This work was done in collaboration with Dr. Richard Keithley who wrote and adapted the MATLAB scripts for the statistical analysis. Cell culture, spheroid growth and IMS experiments were performed in the Hummon Lab. Dr. Keithley assisted in the biological interpretation of some of the statistical data. This chapter was initially published as a research paper in Analytical Methods in 2015.

4.1 Abstract

As imaging mass spectrometry (IMS) has grown in popularity in recent years, the applications of this technique have become increasingly diverse. Currently there is a need for sophisticated data processing strategies that maximize the information gained from large IMS data sets. Traditional two-dimensional heat maps of single ions generated in IMS experiments lack analytical detail, yet manual analysis of multiple peaks across hundreds of pixels within an entire image is time-consuming, tedious and subjective. Here, various chemometric methods were used to analyze data sets
obtained by matrix-assisted laser desorption/ionization (MALDI) IMS of multicellular spheroids. HT-29 colon carcinoma multicellular spheroids are an excellent in vitro model system that mimics the three dimensional morphology of tumors in vivo. These data are especially challenging to process because, while different microenvironments exist, the cells are clonal which can result in strong similarities in the mass spectral profiles within the image. In this proof-of-concept study, a combination of principal component analysis (PCA), clustering methods and linear discriminant analysis was used to identify unique spectral features present in spatially heterogeneous locations within the image. Overall, the combination of these exploratory data analysis tools allowed for the isolation and detection of proteomic changes within IMS data sets in an easy, rapid and unbiased manner. Furthermore, a simplified, non-mathematical introduction to the techniques is provided in addition to full command routines within the MATLAB programming environment, allowing others to easily utilize and adapt this approach.

4.2 Introduction

Matrix-assisted laser desorption/ionization imaging mass spectrometry (MALDI IMS) is becoming an increasingly popular method for a variety of imaging applications ranging from organ sections of animals to disease and disease-free clinical samples [1-4]. MALDI IMS is an unbiased, label-free technique with the ability to detect a variety of molecules including proteins, peptides, lipids, and small molecules over a wide mass range making it the most versatile of the MS imaging techniques [5]. This method of imaging provides extremely rich data sets yielding ion heat maps containing both spatial
and spectral information on the analytes found within the sample. Recent years have seen a remarkable increase in the application of IMS of all types to systems beyond traditional tissue imaging. A growing area of research is the application of IMS to three dimensional cell culture (3DCC) systems [3, 6-8]. These models occupy a distinct niche between conventional two dimensional cell culture systems and animal models [9-11]. Advantages to 3DCC include rapid analysis time, relatively high throughput and experimental flexibility, as compared to animal models [12].

The 3DCC system examined in this study is the HT-29 colon carcinoma spheroid tumor model. When plated under specific growth conditions, cells grow into spheres that can reach up to 1 mm in diameter. At this size, these models develop nutrient, oxygen and transport gradients that generate biologically distinct microenvironments within the spheroids [3, 13]. Importantly, these microenvironments closely resemble different populations of cells found in large tumors in vivo. The outer proliferative region of the spheroid mimics regions of tumors situated closest to blood vessels or other nutrient sources. An intermediate quiescent region is populated by cells that have restricted access to oxygen and nutrients and are alive but no longer actively dividing. The cell culture models also contain an inner necrotic core of dead and dying cells that resemble the population of cells in a tumor furthest from blood vessels with little or no access to oxygen and fresh nutrients. 3DCC models allow researchers a unique opportunity to study the relationships between these heterogeneous cellular populations, and how changes in one region affect the others as well as how these structures are affected by drug or small molecule treatments [5, 6]. The combination of
MALDI IMS and 3DCC takes advantage of the experimental flexibility of 3DCC systems and the wide mass range of MALDI mass spectrometry to yield highly complex and informative data sets. While single-ion heat maps can provide useful information, the application of various data processing approaches is required in order to take full advantage of all the benefits offered by IMS [14, 15]. The use of multivariate statistical analyses on IMS data sets can help highlight otherwise undetected differences among distinct regions and can be especially useful for elucidating changes over time in cell populations as a result of drug treatment or other cell culture perturbation [6, 8, 11, 16, 17]. These methods are especially advantageous when considering the complexity of IMS images (i.e. hundreds of mass spectra containing hundreds of peaks in spatially relevant regions). Multivariate analysis techniques have also been applied to secondary ion mass spectrometry (SIMS) images of rat heart tissue and heterogeneous cellular cultures [8, 17-20].

MALDI IMS presents several unique experimental difficulties that can be corrected using multivariate analyses including matrix interference, apparent mass discrepancies due to sample height differences and non-uniform ionization efficiency [21]. There is no consensus on a universally recognized approach for multivariate analysis of IMS data because each data set presents different challenges. However, each data analysis workflow seeks to accomplish three main goals: the removal of experimental artifacts from the imaging process, reduction of the complex data sets into more computationally manageable formats (without significantly compromising the experimental information contained within) and the successful application of
multivariate analyses to capitalize on what these complex data sets have to offer [22, 23]. In this scheme, each processing step insures that any subsequent multivariate analyses would examine trends present in the data arising from biological variation instead of experimental artifacts. As an example, pixels which contain only noise-related or high-intensity matrix-related peaks are removed because they could bias subsequent multivariate analyses against detecting more subtle changes in protein expression.

Fonville, et al. described the effects of different normalization strategies on subsequent multivariate analysis of MALDI IMS performed on rat brain tissue in the small molecule mass range [24]. They followed the same basic blueprint for the processing of IMS data: isolation and removal of noise-related or matrix-related pixels, transformation of the remaining biologically relevant pixels into a multivariate space and further multivariate analysis of the data to reveal spectral trends identifying different regions of tissue based on patterns in the spectra. In one particular case, after identifying and removing pixels which contained only noise or matrix-related peaks from the data, the authors applied logarithmic scaling to the remaining pixels. This logarithmic transformation reduced the dynamic range of the measurements, allowing the low intensity features in the mass spectra to become more pronounced relative to large peaks. As a result, small, but relevant, changes in the measured intensities the pixels could be factored into the final analysis [25].

In a study by McCombie et al, the authors primarily focused on the combination of principal component analysis (PCA), clustering methods and linear discriminant analysis (LDA) to reveal biologically important regions of mouse brain sections analyzed
by MALDI IMS [26]. This particular approach made use of the iterative application of multiple chemometric techniques to highlight biologically relevant spectral features in the data in an unbiased manner. PCA was first used to reduce the dimensionality of the data set and identify noise and matrix related pixels that were then removed. Next, the authors applied unsupervised clustering methods, including $k$-means and hierarchical clustering analysis (HCA), to find and aggregate groups of pixels that exhibited similarities in protein expression (demonstrated by strong spectral overlap). Finally, LDA was used to generate discriminant spectra that identified specific $m/z$ ratios differentially expressed within distinct regions of the brain tissue.

In addition to the difficulties inherent to MALDI IMS, most 3DCC models lack visible tissue structures that separate biologically distinct regions of interest that further complicates data analysis. Here we integrate several chemometric methods into a single workflow for the processing of MALDI IMS data generated from HT-29 spheroids. First, PCA was used to pinpoint and isolate only the biologically relevant pixels in a MALDI IMS image. Second, after some spectral processing steps, a second iteration of PCA was used to isolate relevant features within the mass spectra of the image. Next, clustering methods including $k$-means and HCA were used to identify and group the relevant pixels into multiple distinct cellular populations based on similarity in the recorded mass spectra. Finally, a combined PCA-LDA approach was used to highlight specific $m/z$ values differentially expressed within the unique cellular populations.
4.3 Results and Discussion

4.3.1 Workflow Overview

The overall research goal was to utilize various multivariate data analysis methods for the identification of spectral features associated with distinct cellular subpopulations within multicellular spheroids. The data shown here is taken from a single HT-29 spheroid slice to show proof-of-concept. An overview of the approach is presented first, followed by a detailed analysis. If needed, the reader is directed to the Appendix C for a theoretical introduction on each of the chemometric techniques.

Figure 4.1 illustrates the statistical workflow for the analysis of a MALDI image of an HT-29 spheroid section. First, raw data files for each pixel within the image were converted into the mzXML format using CompassXport 5.0, imported into the MATLAB programming environment and concatenated into a single data matrix. PCA was then performed on this data matrix containing all the mass spectra. Using a two-dimensional score plot, the spatial distribution pattern of the principal component (PC) scores was then used to isolate physiologically-relevant spectra originating from the HT-29 cells from pixels corresponding to the surrounding gelatin support. The mass spectra of the relevant pixels were then processed to better highlight potentially discriminating spectral features of low abundance, remove baseline shifts, reduce noise, correct for differences in apparent peak m/z, and normalize for differences in absolute inter-pixel intensity. A second iteration of PCA was then performed on the processed data of the relevant pixels to highlight spatially-localized mass spectral features within the data set, further remove noise, and reduce the dimensionality of the data prior to subsequent
data processing steps. Next, cluster analysis was performed on the scores of the relevant PCs to identify and group pixels associated with distinct cellular subpopulations that exhibited spectral similarity within the spheroid section. Finally, PCA-LDA was used to identify specific m/z values associated with proteins differentially expressed within the distinct regions of interest.

Figure 4.1: Summary of data workflow. MALDI IMS data is first imported into the MATLAB programming environment. PCA is then used to isolate relevant pixels within the image. The mass spectra of the isolated pixels are then further processed before a second round of PCA that highlights general spectral features associated within the sample. Clustering analysis is then performed using the scores of the relevant PCs from this second iteration of PCA to reveal the spatial distribution pattern of distinct cellular populations. Finally, the results of both the PCA and clustering of relevant pixels are used as inputs into an LDA routine that identifies specific m/z values selectively associated within each of the distinct cellular microenvironments in the spheroid.
4.3.2 Isolation and Initial Processing of Relevant Pixels

In the first step of the statistical workflow, the recorded spectra were trimmed to a relevant mass range spanning 8-14 kDa and subsequently downsampled. Both of these procedures allow the large IMS data sets to be more computationally manageable for further chemometric processing [22, 27, 28]. While downsampling reduces spectral resolution, it increases processing efficiency by reducing the number of points per spectrum. Trimming the low and high ends of the mass spectra removes noise and ensures that multivariate analyses will be focused only on changes in protein expression within the mass range of interest. Once trimmed, it is necessary to select pixels that arise directly from the sample and discard those that contain mostly noise and matrix-related peaks. In this MALDI IMS image, the majority of pixels are associated with the spheroid cells while others are associated with gelatin and/or matrix artifacts. These latter irrelevant pixels generally exhibited featureless mass spectra for pixels associated with gelatin, mass spectra containing matrix artifacts, or mass spectra that are excessively noisy for accurate analyses as shown in Figure 4.2A. For these reasons, removing the irrelevant pixels made it easier for the subsequent clustering algorithms to identify distinct regions of interest within the spheroid section (vide infra).

A first round of PCA was performed on the trimmed and downsampled data set. PCA is an exploratory analysis tool that deconstructs a multivariate data set into a set of PCs (also termed eigenvectors or loadings) and scores. For MALDI IMS data sets, the loadings highlight independent sources of spectral variation within the data set while the scores describe the importance of each of those spectral trends to each imaged
pixel. The loadings can be thought of as a multivariate version of an independent variable, much the same way $m/z$ is the univariate independent variable within individual mass spectra. The scores act as an intensity factor for a loading at a specific pixel. Pixels with similar mass spectral profiles should have similar PCA score values and aggregate together within a two-dimensional PCA score plot [29]. Based on the clustering pattern of a two-dimensional score plot, the relevant pixels associated with the spheroid cells were identified and retained for further processing while pixels associated with artifacts were discarded (see Appendix C). Figure 4.2B shows the relevant pixels identified using the PCA score plot. Pixels from outside the cell mass or those associated with technical/instrumental artifacts (grey) were successfully separated from the relevant pixels associated with the spheroid cells (white).
Figure 4.2: Isolation of Relevant Pixels and Spectral Processing. A) Example mass spectrum of a pixel discarded after an initial round of PCA on the unprocessed data. B) Isolation of mass spectra from pixels associated with spheroid cells (white) using PCA. Spectra arising from pixels outside of the sample (grey) were discarded. C) The total ion spectrum of all relevant pixels identified in B (white pixels). D) The total ion spectrum after log transformation and baseline removal.

Once relevant pixels were isolated, standard data processing steps for removing MALDI experimental artifacts were applied [22]. A top-hat filter and Gaussian smoothing function were used to remove baseline changes associated with MALDI matrix and reduce noise within the spectra30. Because a time-of-flight (TOF) instrument was used to image the spheroid slice, differences in sample height can cause shifts in the apparent mass of peaks [23]. The m/z alignment of mass spectra is an essential pre-
processing step necessary for TOF data to ensure that regions of interest are identified based on physiological spectral trends associated with protein expression patterns [26]. If spectra are not aligned, subsequent clustering algorithms may falsely identify regions of interest based on topographical artifacts associated with the imaging experiment including sample roughness or sample height differences caused by the spheroid sectioning process. Here, two peaks present in a majority of the mass spectra were used for aligning of the HT-29 spheroid section; in all spectra, the peaks appeared in the approximate ranges of 9120-9200 $m/z$ and 12,520-12,650 $m/z$. In each spectrum, the apparent masses of these two peaks were aligned to the median $m/z$ values recorded for those two peaks across all pixels. The summed spectra across all relevant pixels after baseline subtraction, smoothing and spectral alignment is shown in Figure 4.2C.

The intensities of peaks within an individual mass spectrum can span several orders of magnitude. This large variation makes it difficult for subsequent chemometric methods such as PCA to identify low intensity peaks that offer high discrimination power between different regions of interest within a MALDI image [24]. This issue is typically subverted using variance scaling prior to PCA, a process that allows PCA to assess all $m/z$ values as equally important [31]. Unfortunately, this approach also dramatically increases the influence of $m/z$ values that describe only noise and can thus degrade the quality of the results obtained. Instead of variance scaling, Fonville et al. used an empirical approach whereby all spectra underwent a logarithmic transformation prior to PCA; their approach successfully allowed PCA to easily detect small intensity discriminating peaks without significantly increasing the influence of
noise [24]. Here, the same logarithmic-transformation procedure was used to identify as many discriminating peaks as possible within this data set that showed an unusually high degree of spectral similarity [3]. The final step of spectral pre-processing is normalizing the spectra to the total area under the peaks. This final step removes any remaining variation in the data due to the MALDI measurement process (i.e., uneven ionization of analytes) rather than biological changes in the sample. Figure 4.2D shows a summed spectrum after logarithmic scaling and area normalization.

Figure 4.3 shows data obtained from a second iteration of PCA performed on the relevant processed data from Figure 4.2D. As examples, the loadings from PCs one, three and six are shown on the left side of the figure while the corresponding score heat map for each PC is shown on the right side. Examining PC 1, the loading plot shows several features of the spectra between 10,000 and 12,000 m/z in the top half of the plot, meaning these features are strongly associated with red/yellow-colored pixels towards the center of the spheroid slice. Examples of other spatially-localized features associated with other PCs are also shown in Figure 4.3. Despite the fact that a specific pixel may have a high score value for one particular PC, spectral characteristics identified by other PCs may also contribute significantly to its recorded mass spectrum. As such, these “pseudo-spectra” must be interpreted cautiously because, while they include m/z information of high discriminatory power, one particular PC does not solely describe all relevant information within a specific pixel [32]. Thus, because principal component loadings can sometimes be difficult to interpret further processing is needed.
to identify all of the spectrally-relevant features that discriminate between distinct regions of interest within the spheroid section [33-34].

4.3.3 Cluster Analysis

The purpose of the previous steps in the overall workflow was to isolate potentially discriminating spectral features within the relevant pixels of the MALDI image. Cluster analysis was next used to identify groups of pixels associated with distinct cellular populations within the heterogeneous spheroid structure, assuming distinct
cellular populations exhibited unique mass spectral profiles. Cluster analysis was performed on the scores of the relevant PCs of the processed data. Because cellular populations within the spheroid exhibit the same unique mass spectral signature, all pixels within the MALDI image associated with one particular cellular population should exhibit similarities in their score values for specific PCs. Thus, the score values can be used by cluster analysis to separate all pixels within the MALDI image into distinct cellular populations [35].

Two different methods of cluster analysis were tested here: HCA and k-means clustering. In HCA, interpoint distances between all data elements are calculated, the two closest points are aggregated and treated as a new point, and the process is repeated until all points are joined together; a distance cutoff is then used to segregate the data elements into distinct clusters [36]. In k-means clustering, k number of points are first randomly chosen to be cluster centers and all points are then assigned to the group with the nearest cluster center; the resulting cluster center is calculated and the process is repeated iteratively until convergence [37]. Cluster analysis typically requires that the number of distinct groups be known before the analysis begins. To eliminate subjective bias, the gap statistic was used here to estimate the number of distinct groups for both HCA and k-means clustering as this empirical approach does not require any a priori knowledge of the dataset (Appendix C) [24,38].

The results of the cluster analysis are shown in Figure 4.4. HCA identified two distinct groups of pixels (Figure 4.4A and 4.4B) and k-means identified three distinct groups of pixels (Figure 4.4C and 4.4D). The cluster maps represent distinct groupings of
pixels identified by the clustering algorithms and the traces represent the average of the
processed mass spectra for all the pixels within the particular group. While both
approaches identified distinct populations within the expected concentric pattern, only
$k$-means clustering was able to successfully identify three separate groups associated
with the three well-known microenvironments of 3DCC systems. The combination of the
cluster map in Figure 4.4C and the summed spectra from each region in Figure 4.4D give
an approximation of the similarities and differences in protein expression among the
distinct biological environments. While the PC loadings shown in Figure 4.3 offer some
insight into these relationships, those pseudo-spectra are not directly interpretable.
Figure 4.4 offers directly interpretable mass traces with corresponding cluster maps that
are directly translatable to biological changes within the sample. Figure 4.4D also
highlights the difficulty inherent with the analyses of these systems and the success of
our approach. In general, the same analytes (within this selected $m/z$ range) were
detected in each of the different microenvironments; this behavior is somewhat
expected since all cells are clonal. However, what discriminates these
microenvironments into separate populations is the relative ratios of these analytes.
These subtle differences present a difficult analytical challenge that requires both
finesse and sophistication for proper analysis.
Figure 4.4 Cluster Analysis. A) Regions of interest identified by HCA. B) The average processed spectra of each region in A. The purple and green traces correspond to the average processed spectra of all purple and green pixels, respectively. C) Regions of interest identified by k-means clustering. D) The average processed spectra of each region in C. The purple, green, and blue traces correspond to the average processed spectra of all purple, green, and blue pixels, respectively.

4.3.4 PCA-LDA

The results shown in Figure 4.4D can be used to help identify changes in protein expression patterns within the distinct cellular populations of 3DCC systems. However, a combination of PCA and LDA was used as a final step to clearly visualize m/z values selectively associated within distinct microenvironments within the spheroid.
The field of multivariate discriminant analysis is traditionally used to classify future data observations into different categories based on the similarity to known, well-characterized data. One goal of LDA is to easily visualize the differences amongst groups of data that have previously been clustered into separate groups. To achieve this objective, a series of linear boundaries are calculated and plotted in the coordinate system, segregating a multivariate data space into separate regions associated with different groups of data. These boundaries represent a way to discriminate between different clusters of data observations because future observations are classified into a particular data group depending on which of the separate regions of a multivariate space they are localized.

Specifically, LDA mathematically calculates a boundary that maximizes intergroup separation using a linear combination of independent variables:

\[ \text{Boundary} = AX_1 + BX_2 + CX_3... \]  

(Eq. 4.1)

where \( X_1, X_2, \) and \( X_3 \) are the independent variables and \( A, B, \) and \( C \) are scalar coefficients used to determine direction. In a combined PCA-LDA analysis, the independent variables used to calculate the linear boundaries between groups are the retained PCs. In terms of MALDI IMS data, LDA would normally require that the number of points within a mass spectrum to be less than the number of pixels imaged. By performing PCA prior to LDA, the number of independent variables decreases dramatically from several thousand \( m/z \) values to a few relevant principal components,
enabling the calculation of a unique solution for the linear boundaries by LDA40. Moreover, noise can be removed from the spectra by discarding the non-relevant principal components, thereby improving the linear boundary calculations26.

Figure 4.5 demonstrates the combined PCA-LDA analysis. Figure 4.5A and 4.5B show PCA score plots of the pixels clustered by k-means and the calculated linear boundaries. In both cases, the boundaries successfully discriminate amongst the different spheroid microenvironments (outer versus middle in 4.5A and middle versus core in 4.5B). While a simple line is shown on these two-dimensional score plots for clarity, it is important to note that the linear boundary actually exists as a bisecting hyperplane within the multivariate hyperspace that includes all of the retained PCs (see Appendix C).

Equation 4.1 shows that each of the calculated PCA-LDA discriminating boundaries is simply a linear combination of principal components. Mathematically, this linear combination of principal components can generate a discriminating mass spectrum using the scalar coefficients of the PCA-LDA linear boundary and the PC loadings (Appendix C). These calculated discriminant spectra are shown in Figure 4.5C and 4.5D. The spectral features that are more strongly associated with one group of pixels (specific microenvironment) versus another are on either side of the mean value within a discriminant spectrum. For data that is mean-centered prior to PCA, the mean value within a discriminant spectrum will be zero, greatly improving interpretation because the sign (positive or negative) of the calculated discriminant spectrum at a
specified m/z can be used to determine the selective association of an analyte with a particular group [26].

The calculated discriminant spectra highlight several species that are selectively distributed in the different microenvironments within the spheroid. Perhaps the easiest change to notice first is a large negative peak at approximately 10,890 m/z (labeled i) in the discriminant spectrum shown in Figure 4.5C which indicates that this analyte is more dominant in the middle region of the spheroid as compared to the outer region. This same m/z value also appears as a small negative peak in the discriminant spectrum in Figure 4.5D, indicating that this species is slightly more abundant in the core region as compared to the middle region. Taken together, both discriminant spectra suggest that there exists a species at approximately 10,890 m/z whose abundance increases dramatically from the outer to the middle regions of the spheroid, but only a small amount when moving from the middle region into the core. Figure 4.5E shows a two-dimensional heat map from flexImaging for the species at 10,889 m/z. Indeed, the abundance of this species increases towards the center of the spheroid, with the largest increase being between the outer and the middle regions. This behavior was also shown in the average traces shown in Figure 4.4D. While this difference appears subtle, it is not easily noticeable when examining only the standard ion maps generated by the imaging software and indeed small changes in low abundance proteins could portend more drastic phenotypic changes in the cell populations. The ability to discern these small changes is very important for understanding the biochemical differences in these populations of cells.
Figure 4.5: PCA-LDA Analysis. A and B show score plots of the clustered pixels and the calculated linear boundaries. A) The calculated linear boundary separates the pixels associated with the outer and middle regions of the spheroid. B) The calculated linear boundary separates the pixels associated with the middle and core regions of the spheroid. C) The calculated discriminant spectrum differentiating between the proliferative outer region (shaded blue) and the quiescent middle region (shaded pink). D) The calculated discriminant spectrum differentiating between the quiescent middle region (shaded blue) and the necrotic core region (shaded pink). The lower case Roman numerals highlight several m/z values discussed in the text. E) Single ion heat maps of m/z values identified using the discriminant spectra. The m/z value and its corresponding Roman numeral notation is inset within each two-dimensional heat map.
The discriminant spectra also offer clues as to the biological status of the different groups of cells. For example, at an m/z of 11,005 (ii), there is a moderately-sized peak in the negative region of the outer versus middle discriminant spectrum and a large negative peak in the middle versus core discriminant spectrum. This indicates that the protein at m/z 11,005 is increasing in abundance towards the necrotic center of the spheroid. Both the two-dimensional heat map shown in Figure 4.5E and the average traces shown in Figure 4.4D for 11,005 m/z confirm that the discriminant spectra were correct. Because it is known that the cells in the necrotic core are deprived of nutrients and oxygen, this protein species may be an indicator of dead or dying cells. Conversely, there is a peak at approximately 12,550 m/z (iii) that is positive in both discriminant spectra indicating that this species is more abundant in the middle region as compared to the core region, yet also more abundant in the outer region than the middle region; in other words, the abundance of the species at 12,550 m/z is increasing in abundance outward from the center of the spheroid. Because the cells in the intermediate and outer regions of the spheroid are exposed to oxygen and fresh nutrients, they are generally healthier, with the population in the very outer region still actively dividing. Therefore the protein species at 12,550 m/z could be associated with healthy and/or actively dividing cells. The two-dimensional heat map of this ion shown in Figure 4.5E and the average traces shown in Figure 4.4D again confirm the results obtained by the discriminant spectra.

The average traces in Figure 4.4D show that there is a species at approximately 12,750 m/z whose abundance is similar in the outer and middle regions, but is also
reduced in the necrotic core. There is no peak in the discriminant spectrum of Figure 4.5C at this m/z (iv), indicating that there is no significant difference in the abundance of this species between the outer and middle regions. However, the discriminant spectrum between the middle and core regions shown in Figure 4.5D show a positive peak, indicating that this species is less abundant in the core. Taken together, the discriminant spectra confirm the behavior shown in Figure 4.4D. The two-dimensional heat map at this m/z shown in Figure 4.5E also confirms a moderate abundance of this species in both the outer and middle regions. Interestingly, both discriminant spectra lack a significant peak in the region of 9,155 m/z (v) even though this analyte is prominent throughout the entire image, suggesting that this highly-abundant species offers little in the way of determining differences in cellular populations. Indeed, while this species is of high abundance, the average traces in Figure 4.4D show its average intensity is relatively consistent in all three regions of the spheroid. The two-dimensional heat map in Figure 4.5E also shows that there are areas of high, moderate, and low relative abundance of the 9,155 m/z peak in each of the three regions of the spheroid.

As a final example, the discriminant spectrum in Figure 4.5D shows a negative peak at approximately 11,805 m/z (vi), indicating that this species is more abundant in the necrotic core as compared to the quiescent middle region. This fact is confirmed by the two-dimensional heat map shown in Figure 4.5E. While this species originally appeared insignificant in the original spectra (Figure 4.2) and even as a small shoulder of the more-intense 11,700 m/z peak within the processed data (Figure 4.4), the discriminant spectra were successfully able to identify this analyte as one whose
abundance yields high contrast within different microenvironments of the spheroid. It is likely that a peak of such low relative abundance in the naïve data set could have been overlooked during traditional analysis of this MALDI IMS data set. Moreover, while each of the single ion heat maps shown in Figure 4.5E can be observed in the native data analysis program accompanying the mass spectrometer, discriminant spectra shown in Figure 4.5C and 4.5D allow the user to easily simultaneously visualize how multiple ions relate to one another in each microenvironment of the sample.

There are several caveats in interpreting discriminant spectra. First, while very powerful, discriminant spectra still represent calculated best estimates of spectral features that distinguish among groups of pixels. Second, if groups of points cluster in such a way that a linear boundary does not ideally distinguish between two groups, the discriminant spectrum calculated using this boundary could contain erroneous information [39, 40]. Third, the quality of the results obtained by the PCA-LDA approach is highly dependent on the number of relevant PCs retained. If too few principal components are retained, potentially discriminating features would be left out of the calculated discriminant spectra. Conversely, the quality of the discriminant spectra can also deteriorate if too many noise principal components are retained. If these less important PCs are factored into the PCA-LDA analysis, the noise-containing PCs will show significant overlap in the cluster analysis which would cause the boundary calculated by the algorithm to falsely represent randomly distributed noise as a distinguishing feature of the dataset. Therefore, it is essential to appropriately ascertain
the proper number of relevant PCs to retain in the multivariate model prior to the LDA step.

4.4 Experimental

4.4.1 Cell Culture

HT-29 cells obtained from American Type Culture Collection (ATCC, Manassas, VA) were cultured in 5% CO2 at 37 °C in McCoy’s 5A media (Life Technologies, Grand Island, NY) supplemented with 10% fetal bovine serum (HyClone Laboratories, Logan, UT) and 2 mM L-Glutamine (Invitrogen, San Diego, CA) as described previously3. Cell lines were used within 3 months after receipt or resuscitation of frozen aliquots thawed from liquid nitrogen. The provider assured the authentication of these cell lines by cytogenetic analysis. To induce three-dimensional growth, approximately 6000 cells were seeded in each well of the inner 60 wells in 96-well culture plates (Thermo, Rockford IL) and incubated with complete media changes 10 and 14 days after seeding.

4.4.2 Sectioning and Sample Preparation for MALDI Imaging

Spheroids were harvested according to the gelatin-assisted sectioning method as described previously [3]. Briefly, spheroids were removed from media and washed in 1x PBS (Gibco). A thin layer of gelatin was applied to the bottom of several wells of a 24-well cell culture plate (Thermo) and allowed to cool. After cooling, the washed spheroids were gently placed on the gelatin in the wells and then quickly covered in a layer of
warm, liquid gelatin. The cell culture tray was immediately placed at -80°C until sectioning.

Prior to sectioning, the gelatin blocks were removed from the wells and freeze mounted to sectioning mounts using deionized water and cooled in the cryomicrotome chamber (Leica Biosystems, Nussloch, Germany) set at -30°C. Once samples equilibrated, they were sliced in 14 µm-thick sections. Intact slices were thaw-mounted to a room temperature ITO-coated glass slide (Delta Technologies, Loveland, CO) which was immediately transferred to a desiccator where the slides were stored until MALDI matrix application.

The matrix used was sinapic acid (Sigma) mixed in 50:50 vol/vol HPLC grade water (Burdick and Jackson) w/0.1% trifluoroacetic acid (Sigma) and HPLC grade acetonitrile (Burdick and Jackson, Phoenix, AZ). Matrix was applied by hand using a 1.0 µl GC syringe (Hamilton, Reno, NV) with the aid of a dissecting microscope. Slides were washed twice for 30 seconds in cold acetone (EMD Millipore, Billerica, MA) before matrix application. Washing serves to fix protein and remove lipids and other small organic contaminants [41]. Slides were dried in a vacuum desiccator after each wash. Matrix coverage was inspected with a microscope after each application and was repeated as need. Slides were dried in a vacuum desiccator between each application and stored in the dark in a desiccator until imaging.

4.4.3 MALDI Imaging

MALDI spectra were acquired with a Bruker Autoflex III Smartbeam MALDI-TOF mass spectrometer (Bruker Daltonics, Billerica MA) in linear positive ion mode in the 4-
20 kDa mass range. The laser spot was set to “Ultra” and laser attenuator was manually adjusted prior to each experiment for optimal fluence. Spectra were acquired by flexImaging 2.1 (Bruker) with a raster increment of 70 µm.

4.4.4 Data Analysis

The flexImaging (Bruker) and MATLAB (MathWorks, Natick, MA) software packages were used for all data analyses. flexImaging was used to generate and visualize semi-quantitative, two-dimensional ion density maps. Raw data files were converted into the mzXML format using CompassXport 5.0 (Bruker). MATLAB was then used for all other statistical and chemometric data processing. Specifically, in-house MATLAB command lines were written to open/read the mzXML data, process and transform data (trimming, down-sampling, filtering, aligning, normalization, etc.), perform PCA, and perform PCA-LDA. The MATLAB command lines used for clustering analysis were adapted from Martinez, Martinez, and Solka [42]. A MATLAB function freely available from MATLAB Central’s File Exchange was adapted to isolate data points visually presented within the PCA score plot (Appendix C) [43]. All data sets were mean-centered prior to PCA. Singular value decomposition was used to mathematically perform PCA [44]. MATLAB command lines are available in the Appendix B.

4.5 Conclusions

The advent of mathematical programming environments such as MATLAB allows complex data analysis methodologies to become streamlined and automated. Indeed, excluding the loading of data into the MATLAB programming environment, the
execution of the entire analysis workflow presented here takes less than one hour for the experienced user as compared to days-worth of individual peak analysis within individual pixels. In addition, this semi-automated approach can remove potential bias associated with manual evaluation of data sets.

Chemometric methods including those shown here can dramatically improve the quality of the conclusions obtained from MALDI IMS data sets and their use should certainly continue. In this proof-of-concept demonstration, the combination of PCA, clustering methods and LDA allowed for the easy visualization of differences in protein expression levels across multiple cellular subpopulations within a single MALDI image of a 3DCC section. While it is known that there are distinct regions within the 3DCC models, the cluster maps and discriminant spectra easily identified the spatial relationship among multiple $m/z$ values within the different cellular microenvironments simultaneously. This is important because in most cases a single protein species does not adequately describe all biological variation within a system. Unfortunately, chemometric techniques can appear theoretically obtuse, are sometimes described in the literature with advanced statistical sophistication, and can seem daunting to execute for the inexperienced, yet interested user. To confront these challenges and engender more widespread use among other researchers, we have provided a simplified, non-mathematically intensive introduction to the techniques as well as the MATLAB command lines necessary to perform them (Appendix B).

The authors acknowledge that there are many options for each of the steps presented here and no single workflow is optimal for all MALDI IMS data sets. Because
of the flexibility of the MATLAB programming environment, any IMS data set that can be converted into the .mzXML file format can, with some small adjustments, be processed using the scripts provided in the Appendix B. As noted throughout the paper, many of these steps have previously been used on other IMS samples [24, 26]. The workflow presented in this paper was determined to be optimal for the processing of 3DCC IMS data, a system which presents several unique analytical challenges specific to multicellular spheroid systems based on clonal populations of cells which exhibit a high degree of spectral overlap. With workflows such as the one described here in place, detailed studies of the proteomic effects of various genetic manipulations, drug treatments or co-culture experiments can be undertaken in the future.

4.6 Acknowledgments

Funding for R.B.K. was provided by Roanoke College. Salary support for A.B.H was provided by the Walther Cancer Foundation and this research was supported by National Institutes of Health (1R01GM110406-01) for A.B.H.

4.7 References


152


[40] Y. Tominaga, Chemometrics and Intelligent Laboratory Systems, 1999, 49, 105-115.


[43] MATLAB Central File Exchange

APPENDIX A:

SUPPLEMENTARY INFORMATION FOR CHAPTER 2: SINGLE CELL METABOLIC PROFILING OF TUMOR MIMICS

A.1 Overview of Glycosphingolipid (GSL) Metabolism

![Diagram of GSL metabolism]

Figure A.1: Simplified overview of GSL metabolism in colon cancers.
A.2 Cell Diameters

Average cell diameters from the outer proliferative region, the middle quiescent region, and the pre-necrotic core region were 13.0 ± 0.3 µm, 12.5 ± 0.2 µm, and 12.4 ± 0.3 µm, respectively (N = 50 cells each) Single cells isolated from MCTs treated with chloroquine (25 µM) had diameters of 12.9 ± 0.3 µm, 13.0 ± 0.2 µm, and 12.4 ± 0.2 µm for the outer, middle, and core regions, respectively (N = 50 cells each). Figure A.2 presents a phase-contrast image of a set of isolated cells.

![Phase-contrast micrograph of isolated cells. This image was taken of single cells isolated from the outer region of HCT 116 MCTs. The scale bar represents 25 µm.](image)

Figure A.2: Phase-contrast micrograph of isolated cells. This image was taken of single cells isolated from the outer region of HCT 116 MCTs. The scale bar represents 25 µm.
A.3 CE-LIF Analysis of Cellular Incubation Medium

The HCT 116 MCTS incubation medium containing the fluorescent GSL substrates (GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665, 2 µM each) was analyzed to both assess the purity of the fluorescent GSL substrates and to ensure the HCT 116 MCTS incubation medium did not cause GSL metabolism. A small aliquot (15 µL) of the fluorescent GSL-containing HCT 116 MCTS incubation medium was stored at -80°C at the time of HCT 116 MCTS incubation. A week later, a 5 nM dilution of the HCT 116 incubation medium into the CE running buffer was prepared and analyzed by CE-LIF. The resulting three-color electropherogram is shown in Figure A.3.1.

![Three-Color Electropherogram of the Fluorescent GSL Substrates. The three-color electropherogram represents a 5 nM dilution of the fluorescent GSL substrate-containing incubation medium.](image-url)
A.4 Three-Color Single Cell Electropherograms

All 192 three-color single cell electropherograms were normalized by total fluorescence peak area to correct for differences in fluorescent GSL uptake and laser power differences. Each of resulting area-normalized three-color electropherograms are shown in Figures A.4 for naïve HCT 116 MCTSs and A.5 for treated HCT 116 MCTSs. 33, 32, and 35 cells were electrophoresed from the outer, middle, and core-adjacent regions of naïve HCT 116 MCTSs (respectively) and 29, 30, and 33 cells were electrophoresed from the outer, middle, and core-adjacent regions of treated HCT 116 MCTSs (respectively).
Figure A.4: Electropherograms from Isolated Single Cells of HCT 116 MCTSs A, B, and C show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the outer region of HCT 116 MCTSs. D, E, and F show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the middle region of HCT 116 MCTSs. G, H, and I show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the core-adjacent region of HCT 116 MCTSs. Numerical indices indicate the metabolic product (1-GM1, 2-GM2, 3-GM3, 4-LacCer, 5-GlcCer, and 6-Cer). Red numbers indicate the fluorescent substrate in each channel.
Figure A.5: Electropherograms from Isolated Single Cells of Chloroquine-Treated HCT 116 MCTSs. A, B, and C show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the outer region of HCT 116 MCTSs treated with chloroquine (CQ). D, E, and F show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the middle region of HCT 116 MCTSs treated with CQ. G, H, and I show GM3-BODIPY-FL, GM1-BODIPY-TMR, and LacCer-BODIPY-650/665 metabolism (respectively) in single cells contained within the core-adjacent region of HCT 116 MCTSs treated with CQ. Numerical indices indicate the metabolic product (1-GM1, 2-GM2, 3-GM3, 4-LacCer, 5-GlcCer, and 6-Cer). Red numbers indicate the fluorescent substrate in each channel.
A.5 Effect of Chloroquine on Fluorescent GSL Uptake

Initially, all lasers were used at a power of 10 mW. This laser power was used to generate 25 single cell electropherograms. However, treated cells often gave fluorescent signals outside of the linear range of the avalanche photodiode detector. Laser powers were then decreased to 5 mW, 1 mW, and 1 mW for the 473 nm, 532 nm, and 633 nm lasers, respectively. After decreasing the laser power, 167 additional single cell electropherograms were successfully recorded (30, 25, and 29 cells for the outer, middle, and core-adjacent regions of naïve HCT 116 MCTSs and 29, 25, and 29 cells for the outer, middle, and core-adjacent regions of treated HCT 1116 MCTSs, respectively). Because all fluorescent species measured within a single cell originated from the fluorescent substrate, a sum of the total peak fluorescence within each channel of a single cell electropherogram is indicative of the total amount of fluorescent GSL substrate uptake within the cell. Electropherograms measured for cells taken from the six conditions (outer, middle, and core-adjacent with and without treatment) at the lower laser powers were processed and peak areas were determined as described in the Materials and Methods. While all single cell electropherograms that fall within the linear range of the avalanche photodiode are useful for studying GSL metabolic patterns, only the single cell electropherograms excited with lasers operated at decreased powers were grouped. The total fluorescence was calculated by summing total peak areas of each channel within each single cell electropherogram. Bar graphs of total fluorescence peak area, grouped by fluorophore and location, is shown in Figure A.6. Student’s t-tests
were used to test the hypothesis that chloroquine treatment significantly altered totaled fluorescent GSL uptake.

Figure A.6: Total Fluorescence Peak Areas as a Measure of Fluorescent GSL Uptake in isolated HCT 116 MCTS Cells. A, B, and C show a plot of the total fluorescence signal measured within the BODIPY-FL, BODIPY-TMR, and BODIPY-650/665 channels (respectively) for the 167 single cell electropherograms excited at decreased laser powers. The total fluorescence signal within each trace was calculated by summing the peak areas for each metabolite. Asterisks indicate whether chloroquine (CQ) significantly increased the total fluorescence signal. One, two, and three asterisks indicate p < 0.05, p < 0.01, and p < 0.001, respectively, calculated using a Student’s t-test. The n.s. notation indicates no significant difference. Averages are shown with error bars representing standard error of the mean.

A.6 Percent Relative Peak Areas for all Fluorescent GSL Species

Once peak areas were determined, the percent contribution of each species was quantified as a percent of the total fluorescence within each colored trace of a single cell electropherogram. The percent relative peak areas for each BODIPY-FL, BODIPY-TMR, and BODIPY-650/665 labeled species in each condition are shown in Figures A.7, A.8, and A.9, respectively.
Figure A.7: Average Percent Relative Peak Area for Each Metabolite within the BODIPY-FL Channel as a Function of Location and Pharmacological Treatment. A, B, C, D, E, and F illustrate average percent relative peak areas for GM1-, GM2-, GM3-, LacCer-, GlcCer-, and Cer-BODIPY-FL (respectively). Structures for each metabolite are given above each graph with the ceramide region highlighted blue to indicate the presence of the BODIPY-FL tag. GM3 is underlined to indicate that it was the fluorescent substrate. Averages are shown with error bars representing standard error of the mean.
Figure A.8: Average Percent Relative Peak Area for Each Metabolite within the BODIPY-TMR Channel as a Function of Location and Pharmacological Treatment. A, B, C, D, E, and F illustrate average percent relative peak areas for GM1-, GM2-, GM3-, LacCer-, GlcCer-, and Cer-BODIPY-TMR (respectively). Structures for each metabolite are given above each graph with the ceramide region highlighted green to indicate the presence of the BODIPY-TMR tag. GM1 is underlined to indicate that it was the fluorescent substrate. Averages are shown with error bars representing standard error of the mean.
Figure A.9: Average Percent Relative Peak Area for Each Metabolite within the BODIPY-650/665 Channel as a Function of Location and Pharmacological Treatment. A, B, C, D, E, and F illustrate average percent relative peak areas for GM1-, GM2-, GM3-, LacCer-, GlcCer-, and Cer-BODIPY-650/665 (respectively). Structures for each metabolite are given above each graph with the ceramide region highlighted red to indicate the presence of the BODIPY-650/665 tag. LacCer is underlined to indicate that it was the fluorescent substrate. Averages are shown with error bars representing standard error of the mean.
APPENDIX B:

CHEMOMETRIC ANALYSIS OF MALDI MASS SPECTROMETRIC IMAGES OF THREE DIMENSIONAL CELL CULTURE SYSTEMS: MATLAB PROCESSING USER GUIDE

The following is an instructional manual for the processing MALDI IMS datasets. As expected, this is specific for our work, but the commands can be adapted for other uses. Some general guidance is given on how to execute specific commands, but users should read comments in the .m files provided for a thorough review. The authors assume the reader has read through Appendix C and has a general knowledge of MATLAB. If MATLAB is new to the user, several general texts are available. While it describes an older version, the authors recommend “Getting Started with MATLAB: A Quick Introduction for Scientists and Engineers” by Rudra Pratap.

We, the authors, provide these instructional materials in good faith and we believe them to be free of errors to the best of our knowledge. This document was written for clarity and ease of execution. As such, in some cases the command lines may not include the most direct way to accomplish a desired task, but (as the authors feel) they are easier for the novice user to follow and execute. The authors are in no way liable for any damages (e.g. software inoperability, data loss, etc.) associated with the use of this material.

While some workarounds have been attempted in some cases, several toolboxes are required including the Signal Processing, Bioinformatics, Image Processing, and
Statistic Toolboxes. It is also assumed that all of the following .m files are located within the MATLAB directory:

- open_read_interp.m
- mzPCA_raw.m
- selectdata.m
- alignspectra.m
- mzPCA_processed.m
- gapstatkmeans.m
- kmeansclustering.m
- PCALDA.m
- gaussfilterdata.m
- plotandselect.m
- subdatapreprocessing.m
- align_and_normalize.m
- image_gen.m
- gapstatHCA.m
- HCAclustering.m
- PCALDA_higher_order.m

Redistribution and use in source and binary forms, with or without modification, are permitted provided that redistributions in binary form must reproduce the copyright notice, this list of conditions and the following disclaimer in the documentation and/or other materials provided with the distribution. This software is provided by the copyright holders and contributors "as is" and any express or implied warranties, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose are disclaimed. In no event shall the copyright owner or contributors be liable for any direct, indirect, incidental, special, exemplary, or consequential damages (including, but not limited to, procurement of substitute goods or services; loss of use, data, or profits; or business interruption) however caused and on any theory of
liability, whether in contract, strict liability, or tort (including negligence or otherwise) arising in any way out of the use of this software, even if advised of the possibility of such damage.

B.1 Part One: PCA of Relevant Pixels

B.1.1 Load Data (Optional)

Summary: Use the open_read_interp command to load the mzXML files into the MATLAB programming environment. Because each pixel contains slight variations in the absolute m/z values recorded (e.g. 1000.97 vs 1100.05), all spectra are interpolated such that they all share the same mass axis.

CompassXport 5.0 (freely available on Bruker’s website) was used here to generate mzXML files for each pixel from the raw data. As such, the ability to open and read in data using our MATLAB commands is selective for this method. However, users who have an alternate approach to importing their data in MATLAB can skip this step or adapt the open_read_interp.m file solely for the purposes of interpolating their data. If taken from somewhere else, going forward users must have the following: a matrix containing XY coordinates for all mass spectra, an m/z axis, and mass spectral data that are interpolated such that all spectra share the same mass axis. See the open_read_interp command for more specific details on formatting for each of these matrices (rows vs columns).

If using our approach, all mzXML files must be located within a specific directory that has no other files within it. Any files that were under 100 kB were deleted prior to
executing this command as these files contained incomplete spectral information associated with irrelevant pixels (e.g. an incomplete spectrum recorded by the mass spectrometer with only 20 points).

Command:

[mzaxis, interpdata, coords] = open_read_interp(directory, dataname, lowmzlimit, highmzlimit, spacing);

The inputs include the name of the directory where the mzXML files are located, a data name that the workspace will be saved as when the file loading is finished, and a way for the user to set the lower limit, upper limit, and spectral m/z resolution of the loaded spectra. The outputs contain the m/z axis, the interpolated spectra for all pixels, and the XY coordinates for all pixels.

Example:

[mzaxis, interpdata, coords] = open_read_interp('C:\Users\Hummon Lab\Desktop\ctrl_sp1_2', 'Ctrl1_2_raw.mat', 7000, 14000, 1);

This command would look for files within a folder named ctrl_sp1_2 on the Desktop, interpolate all mass spectra such that all data spans from 7000 to 14000 m/z at unit resolution, and would save the spectra as a file called Ctrl1_2_raw.mat within the same directory. The three outputs would be named mzaxis, interpdata, and coords for the m/z axis, the interpolated data, and the XY coordinates (respectively). The user should be aware that this command can take a significant time to execute, depending on the amount of data to be imported.
B.1.2 Trim Data (Optional)

Summary: You may trim the ends of the mass spectra to isolate only the relevant $m/z$ range. To do this without accidentally altering the original data (in case you need it), we make a copy and trim the copy. You must trim both the mzaxis and interpdata. The trimming is done on a point index basis, not on an absolute $m/z$ basis.

Commands/Example:

```
[trimmed]=interpdata;
[mzaxis_t]=mzaxis;
trimmed(:,1:1000)=[];
mzaxis_t(1:1000)=[];
```

The above commands duplicate both interpdata and mzaxis as well as remove the first 1000 points from each. You should look through the mzaxis vector prior to executing this command to see the range of $m/z$ values that are removed. Using the data above that has unit mass resolution, these commands would remove masses 7000 to 7999. The remaining data would span from $m/z$ 8,000 to 14,000 (inclusive) giving 6001 data points per spectrum. This procedure can also be adapted to remove irrelevant $m/z$ values on the upper end of the dataset (high mass range). While our example data below in subsequent steps does not include the following step, as an example a user could type:

```
trimmed(:,5502:end)=[];
mzaxis_t(5502:end)=[];
```
The above commands would remove the last 499 data points from the mass spectra. Using the data above that has unit mass resolution, these commands would remove masses 13,501 to 14000, giving 5501 data points per spectrum. Again, this was not done for the subsequent data analyses, but the commands provide an example of how to accomplish the trimming task on the upper end of data.

B.1.3 Gaussian Smooth Data (Optional)

Summary: The following commands will utilize a Gaussian smooth to remove noise from mass spectra. If used, this step must be performed prior to downsampling to prevent errors associated with peak shifts. Specifically, each of the mass spectra within the image are convolved with a Gaussian function to remove noise. After this command, a graph will pop up which will allow the user to assess the quality of the smoothing step. The command can be repeated multiple times with different parameters (overwriting the previous iteration) to allow adjustment if necessary.

Command:

\[
[\text{filtered}]= \text{gaussfilterdata}(\text{mzaxis}, \text{dataraw}, \text{sigma}, \text{pixelnum});
\]

The inputs include the mass axis (may be trimmed), the raw data (again may be trimmed), a scalar number that helps defines the width of the Gaussian smoothing function, and a scalar number used to identify a specific pixel. This pixel number will be the spectrum that is viewed to assess the quality of the data processing.

Example:

\[
[\text{trimmed\_f}]=\text{gaussfilterdata}(\text{mzaxis\_t}, \text{trimmed}, 3, 88);
\]
Here, this command would smooth the data with a Gaussian that has a standard deviation of 3 \textit{m/z} (because we have unit resolution), and the spectrum corresponding to the 88th pixel will be viewed to assess the quality of the Gaussian smoothing. An example graph produced by this command is shown on the left in Figure B.1; an enlargement of one area is also presented on the right. Note that the two traces shown are the original data (blue) and the smoothed data (red). As expected with a Gaussian smooth, there is a balance between noise removal and peak amplitude suppression. The user should use what is best suited for their particular data set.

![Figure B.1: Example comparing two variations on using Gaussian smoothing to remove noise from mass spectrometry data. Left graph shows an entire summed spectrum while the graph on the right is a close up of the region between 9000 \textit{m/z} and 9900 \textit{m/z} units. In both examples the blue trace indicates the original, non-smoothed data while the red trace represents the summed spectra after the smoothing algorithm has been applied.](image-url)
Downsample Data (Optional)

Summary: You may choose to downsample the mass spectra to enable faster processing. If the computer lacks sufficient processing power, the subsequent PCA command sometimes will not execute on large datasets. In this case, MATLAB would return an out of memory error when trying to execute. There is no absolute rule as to when this can occur and depends only on the computer being used. As an alternative, this step could be completed after the first iteration of PCA (provided your computer has sufficient processing power to perform PCA on large datasets). The user would then just have to adapt the names of the matrices used in the commands. This step takes advantage of MATLAB’s downsample command.

Commands/Example:

```matlab
[mzaxis_d]=downsample(mzaxis_t,10);
for x=1:size(trimmed_f,1)
    trimmed_d(x,:)=downsample(trimmed_f(x,:),10);
end
clear x
```

Alternatively, if one lacks the Signal Processing toolbox, one could simply use indexing to extract every nth point as shown below:

Alternate Commands/Example:

```matlab
[mzaxis_d]=mzaxis_t(1:10:end);
for x=1:size(trimmed_f,1)
```
trimmed_d(x,:)=trimmed_f(x,1:10:end);

eend

clear x

In both cases, the inputs include the mass axis (which you could have trimmed or not), the raw spectra (could be raw, trimmed and/or smoothed), and a scalar that represents the rate at which you wish to downsample which is typed in the 1st and 3rd lines of the above commands. Here, because the number 10 was used, these commands would downsample the trimmed smoothed data at every 10th point. Therefore, the downsampled range would span 601 points instead of 6001 (it keeps the extra point on the end as is). For the commands used here, a downsampled mass axis (mzaxis_d) and a downsampled, trimmed smoothed data set (trimmed_d) would be generated. As another example, if you wanted to down-sample by 5 to get every 5th point, replace the number 10 in the 1st and 3rd lines of the above commands with the number 5. You can type something similar to the commands on the next page to visualize the down-sampling result (total ion spectra are shown below in Figure B.2 before and after downsampling).

Commands/Example:

```matlab
Figure;plot(mzaxis_t,sum(trimmed_f),'LineWidth',2)

hold

plot(mzaxis_d,sum(trimmed_d),r','LineWidth',2)

legend('Before','After')
```
set(gca,'FontSize',16)

set(gcf, 'color', 'white');

xlabel('Mass to Charge Ratio (m/z)', 'FontSize', 16)

ylabel('Intensity', 'FontSize', 16)

Figure B.2: Comparison of summed spectra before and after downsampling is applied.

While downsampling is sometimes necessary, it can reduce your resolution and cause slight errors in the location of peaks (depending on how sharp peaks are, downsampling can catch the edge of a peak, rather than its maximum). Therefore, if comparing the results of the discriminant spectra to those of a commercial mass spectrometer programming package, you should allow an error of +2 of the resolution of newly-down-sampled spectra.
B.1.4 PCA on All Pixels

Summary: This command will generate the scores that are used in the next step to select relevant pixels.

Command:

\[ \text{[Aproj\_raw]} = \text{mzPCA\_raw(rawdata)}; \]

The input is the raw data (may be raw data or trimmed and/or Gaussian-smoothed and/or downsampled data). The output Aproj\_raw contains the scores of the pixels along the first 20 (default) principal components.

Example:

\[ \text{[Aproj\_raw]} = \text{mzPCA\_raw(trimmed\_d)}; \]

Here, this command would perform PCA on the trimmed, Gaussian smoothed, and downsampled raw spectra.

B.1.5 Isolation of Relevant Pixels

Summary: This command will allow users to select and isolate a subset of specific pixels from a PCA score plot of the raw data. When the first plot comes up, immediately left-click and hold down the mouse to “lasso” the points. However, if the plot is too small, you can drag the edges or maximize the window to make it larger as long as you do not click inside of the graph. Make a complete enclosure before letting go of the mouse. Highlighted data points will change color from blue to red and be encompassed in a yellow color as you lasso. When you are finished, a second graph will pop up which shows the XY coordinates of the pixels that you have selected. You may repeat the
process as many times as you need to such that the proper data is isolated. Irrelevant pixels associated solely with matrix, noise, etc. should aggregate outside of the range of the other pixels. If you choose, you may downsample the data after this command executes with subdataraw as an input.

Command:

```
[subcoords,subdataraw]=plotandselect(Aproj_raw,coords,rawdata);
```

The inputs include the PC scores of the raw data, the XY coordinates of the data, and the raw data (may be raw data or trimmed and/or Gaussian smoothed and/or downsampling data). The outputs include the subset of XY coordinates and mass spectral data corresponding of the pixels the user selected.

Example:

```
[subcoords,subdataraw]=plotandselect(Aproj_raw,coords,trimmed_d_d);
```

Here, the user would isolate a selected pixel subset, extracting the corresponding XY coordinates and the trimmed, smoothed, downsampling data. The data subsets generated would be named subcoords and subdataraw for the XY coordinates and the selected data, respectively.

When isolating relevant pixels the results during the execution of this command, the first graph should resemble something shown in Figure B.3 on the left. Note: Other data will have a different distribution of points unique to that particular dataset. After selection of the relevant pixels, another graph will appear that shows the XY coordinates of the isolated subset of pixels as shown below in Figure B.3 on the bottom right. White
pixels are the ones that the user selected and grey pixels are those that were discarded.
The black pixels were never imaged by the MALDI. The numbers on the axes of this second graph correspond to X and Y coordinates of the pixels. If the user makes a mistake or doesn’t agree with the assignments, the same command line can be repeated, overwriting the data generated by previous iterations.

Figure B.3: The left-hand frame shows the isolation command which allows the user to select which pixels to keep for the next steps of the processing (surrounded by yellow circle). The right hand side shows a representative image of the spheroid being processed in the MATLAB environment. Area shaded in white corresponds to the pixels highlighted in yellow in the selection step (pictured left). Gray pixels are discarded from the data set and not processed any further in the workflow.

B.1.6 Pre-Processing of Selected Subset of Data

Summary: This command will log-transform and top hat filter the subset of mass spectra selected from the previous step. During this command, a graph will pop up which will allow the user to assess the quality of the processing steps. The command can be repeated multiple times with different parameters (overwriting the previous iteration) to allow adjustment if necessary.
Command:

\[\text{[processed]} = \text{subdatapreprocessing(mzaxis, subdataraw, tophat, pixelnum)};\]

The inputs include the mass axis (may be trimmed and/or downsampled), the subset of data that includes only the relevant pixels (again may be trimmed and/or Gaussian-smoothed and/or downsampled), a scalar number that defines the top hat filter, and a scalar number used to identify a specific pixel. This pixel number will be the spectrum that is viewed to assess the quality of the data processing (Figure B.4).

Example:

\[\text{[processed\_d]} = \text{subdatapreprocessing(mzaxis\_d, subdataraw, 250, 75)};\]

Here, this command would log-transform the data, use a top hat filter of 250 (removes baseline changes greater than 250 data points wide), and the spectrum corresponding to the 75th pixel will be viewed to assess the quality of the top hat filter.
B.1.7 Align and Area-Normalize

Summary: This command will align all data such that all peaks should occur at the same $m/z$ and normalize all spectra to unit area to correct for pixel-to-pixel differences in absolute intensity of the recorded mass spectra. To accomplish the aligning, two peaks are identified by the user that exist within all pixels of the data. For each of the two peaks, the algorithm will first determine the $m/z$ value of the peak as it occurs within each spectrum. Next, the median $m/z$ values of the two peaks are identified. Third, all spectra are aligned such that the two peaks occur at the same $m/z$ ratios in all spectra (the median $m/z$ values previously identified). The user can also determine how to best normalize the data: area, height, or no normalization. A graph
will also be displayed to show the final resulting data, an example of which is shown in Figure B.5.

Command:

```
[finaldata]=align_and_normalize(mzaxis,processed,P1low,P1high,
P2low,P2high,type);
```

The inputs include the mass axis (may be trimmed and/or downsampled), processed spectra of the relevant pixel subset, a range of \( m/z \) values that encompass one peak of lower \( m/z \) in all spectra, a range of \( m/z \) values that encompass one peak of higher \( m/z \) in all spectra, and the specified normalization type. The output will consist of aligned, area-normalized, processed data of the relevant pixel subset. The user must first plot the data to identify the mass spectral ranges of the two peaks used for aligning. The command below can be used (although the plot will not be labeled as it is below in Figure B.1.5).

Command:

```
figure;plot(mzaxis_d,processed_d);
```
Figure B.5: Top graph shows all of the spectra plotted on top of one another over the entire mass range to be analyzed. The bottom panels demonstrate zooming in on specific peaks that occur in all of the spectra. When the “align_and_normalize” command is performed all of the spectra will be aligned to the center of the two peaks shown in the bottom panel.

Zooming in shows that a lower m/z peak ranges between 9120-9200 m/z among all pixels and a higher m/z peak ranges between 12520-12650 m/z among all pixels.

Using this information, we may write the following command:

```
[finaldata_d]=align_and_normalize(mzaxis_d,processed_d,9120,9200,12520,12650,'area');
```
Here, this command would align all spectra such that the peaks that occur between
9120-9200 m/z and 12520-12650 m/z are aligned to the median m/z value for those
peaks across all pixels. The output data are also area-normalized. When this command
has finished executing, a graph will display to show the final result of the aligning and
normalization algorithm as shown in Figure B.6. In this case, the intensity axis represents
normalized intensity because area-normalization was used.

![Graph](image)

**Figure B.6:** Graph displayed following proper execution of the
“align_and_normalize” command.

**B.1.8 Perform PCA on the Processed Data of Relevant Pixels**

**Summary:** This command will perform PCA on the processed data of the relevant
pixels. A Scree plot will also be generated for the user to estimate the proper number of
principal components to retain.

**Command:**

```
[Aproj_final,Vc_final]=mzPCA_processed(finaldata);
```
The input includes the final processed data. The outputs include the scores of each principal component along each pixel (Aproj_final) and the principal components (Vc_final).

Example:

\[[Aproj\_final\_d, Vc\_final\_d] = mzPCA\_processed(finaldata\_d);\]

This command generates the scores and principal components of the final processed data. Upon execution, the Scree plot is generated. This command automatically mean-centers the data. Any data set that is mean-centered prior PCA will cause a slight defect in the appearance of the Scree graph. Statistically, mean-centering removes one degree of freedom from the data being analyzed because all spectra were used to calculate a spectral average. Normally, the total number of principal components (relevant and noise principal components) equals the number of data elements (e.g. pixels) within a data set. However, by subtracting the spectral average prior to PCA, one principal component has already been removed by the user. Since principal components are calculated in order of decreasing importance, mean-centering typically causes the last principal component calculated to contain zero variance. The logarithmic-transformation of the variance captured by each principal component then gives an error/outlying value for the last principal component that can be ignored. Zoom into the first part of the graph to estimate the proper number of principal components to retain.

An example is shown in Figure B.7:
Figure B.7: Example of a Scree plot zoomed into the first region. Arrows indicate “elbow regions” representing changes in the variance described by each principal component. This method can be used to estimate the number of principal components that contain actual experimental variation and will be retained in the subsequent processing steps.

From the above graph, three “elbows” can be seen (8, 10, and 19). These points are taken to be the maximum number of relevant PCs, but many chemometric researchers (including these authors) remove one PC from the elbow; the elbow maximum is considered to contain the first noise PC so many users deduct one PC from the elbow to deduce the proper number of PCs to retain. In this case, there are three options: 7, 9, and 18 (red arrows above).

Unfortunately, this phenomenon of multiple breaks is sometimes seen with the Scree graph where the proper number of principal components to retain is unclear. It is possible that 18 principal components describe all relevant sources of information. However, it is also possible that only 7 or 9 principal components describe relevant
information while all higher numbered principal components represent heteroscedastic (non-random) noise. One way to determine which is best is to plot the principal components (Section B.1.10) and visualize the spectral trends. Alternatively, the user may opt to choose fewer numbers of principal components moving forward in the process as the quality of discriminant spectra is highly dependent on noise. While it is possible that some information would be left out of the discriminant spectrum using fewer principal components, the peaks that are present in the discriminant spectrum will be of better quality (easier to visualize and have a higher signal-to-noise ratio).

B.1.9 Plot the Principal Components and Score Images (Optional)

Summary: This command will allow users to visualize the principal components produced by the previous commands and to generate a spatially-localized spectral trend “heat map”. Two graphs will pop up for the users, one displaying the principal component and the other presenting the heat map. Users have several options on how to best visualize the data as described in the comments section within the image_gen.m command file.

Command:

```plaintext
image_gen(Aproj_final,Vc_final,subcoords,mzaxis,PCnum,negate, cleangraph);
```

The inputs include the scores and principal components of the processed data (from the previous mzPCA_processed command), the subset of XY coordinates selected with the plotandselect command, the mass axis of your data, the principal component number you wish to visualize, and a set of values as to whether or not the user wishes to negate
the principal component for easier viewing and to remove ancillary features from the heat map graph for better viewing (T for yes, F for no for the latter two inputs). There are no outputs.

Example:

```matlab
image_gen(Aproj_final_d,Vc_final_d,subcoords,mzaxis_d,1,'F','F');
```

The above command will plot the spectral trend identified within the first principal component without negation and will generate a labeled heat map as shown below in Figure B.8:

![Figure B.8](image)

Figure B.8: The spectral trend is shown on the left and the heat map of the spectral trend is on the right. The color scale represents the magnitude of the scores of this particular principal component for each pixel; large positive values indicate a positive correlation with a particular pixel and large negative values indicate a negative correlation. Note: In the heat map, the deepest darkest blue areas indicate pixels that were not isolated by the plotandselect command.
Because principal components are calculated without direction, it is perfectly reasonable to invert (multiply by -1) the principal component. To do this and to generate a heat map that is “cleaner” for easier exporting, the following command can be used. Note that negating the principal component also inverts the image as shown in Figure B.9.

Command:

```matlab
image_gen(Aproj_final,Vc_final,subcoords,mzaxis_d,1,'T','T');
```

Figure B.9: Example of inverting the principal component. The heat map also inverts. This is an example of the “cleangraph” output feature of this command.

B.2 Part Two: Clustering

B.2.1 Estimation of Cluster Number (Optional)

Summary: The following commands estimate the number of clusters that exist within a dataset using the gap statistic. If the user knows the proper number of distinct groups beforehand, this step may be omitted. Because both k-means and hierarchical
clustering analysis (HCA) can be used in the estimation of the gap statistic, both sets of commands are used here. Choose the one that is the most appropriate. In either case, the gap statistic can generate two different types of reference distributions needed for its calculation. When finished, a graph will be generated that mimics Figure 5.10 in Martinez and Martinez (Figure B.10)[4].

For \( k \)-means use:

\[
[khat]=\text{gapstatkmeans}(A\text{proj\_final},\text{PCs},\text{type});
\]

The inputs to this command include the scores generated by the mzPCA_processed command on the processed data, the proper number of relevant principal components, and a choice of type for the two types of reference distributions that are used for the gap statistic[4]. The inputs for type are ‘uniform’ and ‘pc’. The authors used the pc method as it will preserve the underlying distribution of the data set. The output \( khat \) indicates an estimate of the number of distinct groups of pixels within the data set.

For HCA use:

\[
[khat]=\text{gapstatHCA}(A\text{proj\_final},\text{PCs},\text{type});
\]

The inputs have the same meaning as above.

Example:

\[
[khat]=\text{gapstatkmeans}(A\text{proj\_final\_d},7,'pc');
\]

This command performs the gap statistic calculation using \( k \)-means to generate clusters. Only 7 of the 20 principal components generated by the mzPCA_processed command are used for cluster estimation. Also, this command uses the pc method for a reference data set.
An HCA example is shown below:

\[ khat = \text{gapstatHCA}(A\text{proj}\_\text{final}_d, 7, 'pc'); \]

This command performs the gap statistic calculation using HCA. Only 7 of the 20 principal components generated by the \text{mzPCA\_processed} command are used for cluster estimation. Again, this command uses the pc method for a reference data set.

![Graph](image.png)

Figure B.10: A graph will be generated that displays dispersion as a function of cluster number for the reference and test data sets as shown above. See [4] for more information.

### B.2.2 Cluster Analysis

**Summary:** The following commands perform cluster analysis on the scores of the relevant PCs of the processed data by either \textit{k}-means or HCA. Choose the desired approach that is the most appropriate. Both methods can generate different clustering results as these methods cluster data points in a different manner. Regardless of the choice of clustering algorithm, two graphs are displayed to the user: average traces of all processed pixel spectra for a particular cluster type and a map highlighting the spatial localization of the pixels.
For \textit{k}-means use:

\[
[\text{allmeanspectra, idx}]=\text{kmeansclustering}(\text{Aproj\_final, PCs, khat, subcoords, mzaxis, finaldata});
\]

The inputs to this command include the scores generated by the \textit{mzPCA}\_processed command on the processed data, the proper number of relevant principal components, an estimate of the number of clusters that exist within the data set (taken from the gap statistic or a priori knowledge), the XY coordinates of the relevant pixels, the \textit{m/z} mass axis, and the final normalized, aligned, processed data (from the \textit{align}\_and\_normalize command). The outputs consist of averages of all the log-transformed processed spectra for each cluster identified and a set of indices identifying which pixel belongs with which cluster.

For HCA use:

\[
[\text{allmeanspectra, idx}]=\text{HCAclustering}(\text{Aproj\_final, PCs, khat, subcoords, mzaxis, finaldata});
\]

The inputs and outputs have the same meaning as above.

A \textit{k}-means example is shown below:

\[
[\text{allmeanspectrakm, idxkm}]=\text{kmeansclustering}(\text{Aproj\_final\_d, 7, 3, subcoords, mzaxis\_d, finaldata\_d});
\]

This command performs \textit{k}-means clustering using the scores of the first seven principal components of the aligned and processed data to identify three groups of pixels. All other inputs have their same meaning as previously denoted.

An HCA example is shown below:

\[
[\text{allmeanspectraHCA, idxHCA}]=\text{HCAclustering}(\text{Aproj\_final\_d, 7, 3, subcoords, mzaxis\_d, finaldata\_d});
\]
This command performs HCA clustering using the scores of the first seven principal components of the aligned and processed data to identify three groups of pixels. All other inputs have their same meaning as previously denoted.

Do not worry about cluster order (1 vs 2 vs 3) as that is arbitrarily determined by the clustering algorithm. Note that the cluster order may change as the command is repeated, but the groupings of points should remain consistent. Only in the case where three clusters are found by the algorithms will the color of the average trace correspond with the color on the spatial map (e.g. the average spectra for all green pixels within the right graph is shown as the green trace in the left graph of Figure B.11). In all other cases, users should use the graph legends to interpret the data. The color scheme for the spatial map will also change if the number of clusters does not equal three.

Figure B.11: The graph on the left displays the average processed spectra for pixels of a specific group. The legend is numbered data1, data2, data3... to identify the cluster number associated with each average trace. The graph on the right displays an XY coordinate map of the spatial location of each type of pixel. Note: The color bar identifies which region corresponds to which cluster number. In this example, blue identifies the pixels associated with cluster number 1, green identifies the pixels associated with cluster number 2, red identifies the pixels associated with cluster number 3, and black represents irrelevant pixels not lassoed from the plotandselect command and those not originally imaged (0th cluster).
Before moving on to the PCALDA step, the user must remember/record/acknowledge which cluster number corresponds to which spatial region. These numeric indices will be used as inputs for the PCALDA command. There is no way to record this in MATLAB and re-running the command line could possibly associate the different regions with a new cluster number. Here, the outer region of the spheroid is represented by cluster number 3, the middle region of the spheroid is represented by cluster number 2 and the core region of the spheroid is represented by cluster number 1. Again, the assignments are arbitrary.

B.3 Part Three: PCA-LDA

B.3.1 Performing PCA-LDA and Obtaining a Discriminant Spectrum

Summary: The following command utilizes the relevant principal component scores of the normalized, aligned, processed data and the indices obtained by the clustering algorithm to obtain a discriminant spectrum that segregates between two groups of data. This command generates two graphs: a graph of the discriminant spectrum and a score plot of the first two principal components with the linear discriminant. As a note to the user, the borders/lines drawn on the discriminant spectra assume that the data has been mean-centered (which is done by default by the mzPCAprocessed command).

Command:

\[
[\text{discrimspectra}]=\text{PCALDA}(\text{mzaxis,Aproj\_final,Vc\_final,PCs,\text{idx,clustera,clusterb,graphtype}});
\]
The inputs to this command include the mass axis, the scores and principal components generated by the mzPCA_processed command, the proper number of relevant principal components, the indices of cluster number for all pixels from the clustering command, the two cluster numbers that the user wishes to segregate against, and a user input as to the format of the graph for the discriminant spectrum (singletrace or colored). The output is an array containing the discriminant spectrum. Note: Positive values in the discriminant spectrum are associated more strongly with cluster a and negative values are associated more strongly with cluster b. Again, knowing which cluster number is which is essential to properly attributing spectral trends to the appropriate group of pixels.

Example:

\[\text{[discrimspectra}_o\text{ vs m]}=\text{PCALDA}(\text{mzaxis}_d,\text{Aproj}_\text{final}_d,\text{Vc}_\text{fin al}_d,7,\text{idxkm},3,2,'\text{colored}')\];

This command generates a discriminant spectrum using the first seven principal components of the processed relevant pixel subset, comparing cluster number 3 versus cluster number 2. The resulting discriminant spectrum will give a colored, shaded graph. Using the data and associations from step 11, positive values in the discriminant spectrum would yield spectral trends more strongly associated with the outer region (cluster 3) of the spheroid and negative values would yield spectral trends more strongly associated with the middle region of the spheroid (cluster 2). The graphical outputs of the above command are shown below in Figure B.12:
Figure B.12: The graph on the left shows a cluster plot of the principal component scores for each group identified. The dark line is the linear discriminant/boundary calculated by PCALDA. Note that its position successfully segregates the two clusters of data. The graph on the right shows the discriminant spectrum. When ‘colored’ mode is used, positive values will be enclosed within an area of blue while negative values will be enclosed in red. The line through the center marks zero (this will only work for mean-centered data).

Continuing with the above example:

```matlab
[discrimspectra_m_vs_c]=PCALDA(mzaxis_d,Aproj_final_d,Vc_final_d,7,idxkm,2,1,'singletrace');
```

This command will generate a discriminant spectrum that differentiates between the middle (cluster 2) and core (cluster 1) regions. Using the data and associations from step 11, positive values in this discriminant spectrum would yield spectral trends more strongly associated with the middle region (cluster 2) of the spheroid and negative values would yield spectral trends more strongly associated with the core region of the spheroid (cluster 1). The graphical outputs of the above command are shown below in Figure B.13.
Figure B.13: Graphical outputs of the “discrimspectra” command. As expected, the linear discriminant shifts to separate groups 2 and 1. Note that the discriminant spectrum is not shaded because the ‘singletrace’ distinction was used in the command line. The dotted line is used to mark the zero position. In either type of graph, the user may wish to use the zoom tool to identify peaks within the calculated discriminant spectra.

While all relevant principal components are used in the calculation of the discriminant spectrum, by default, the scatterplot generated by the PCALDA command only shows the first two principal components. These may or may not highlight the appropriate PCs that ideally segregate two data sets. An adapted PCALDA command can be used to plot other PCs against one another:

\[ \text{[discrimspectra]} = \text{PCALDA\_higher\_order(mzaxis,Aproj\_final,Vc\_final,PCs,idx,clustera,clusterb,graphtype,PCx,PCy);} \]

Similar to the other PCALDA command, the inputs contain the mass axis, the pixel scores, the relevant principal components, the number of relevant principal components, the cluster indices of all pixels, the cluster numbers of the two groups you wish to discriminate against, and the type of discriminant spectrum graph desired. Two new inputs are now included: the two PCs you wish to plot against one another. You
may choose any values for PCx and PCy up to a certain maximum (the number of relevant principal components). Remember, different clusters of points may share similar spectral characteristics, so you will likely not have segregation of the different clusters in all dimensions. In this case, a plot would be generated of PC3 (y-axis) vs PC2 (x-axis). Using the data and matrix names above, an example command is given by:

\[
\text{[discrimspectra]} = \text{PCALDA\_higher\_order(mzaxis\_d,Aproj\_final\_d,Vc\_final\_d,7,idxkm,2,1,'colored',2,3)};
\]

B.4 References


APPENDIX C:
CHEMOMETRIC ANALYSIS OF MALDI MASS SPECTROMETRIC IMAGES OF THREE-DIMENSIONAL CELL CULTURE SYSTEMS

The material presented in the below first includes an introduction to the theory of several chemometric methods, followed by a full discussion of the statistical workflow used within the main paper. The discussion below is presented on a “need to know” basis to those without significant knowledge of the field and requires only a basic knowledge of introductory statistics. Therefore, occasionally details/specifics of the techniques are omitted for clarity and oversimplifications are made to improve flow. There are several introductory texts cited within this material that can be used by readers for further information and study.

C.1 Principal Component Analysis

Data sets that record one independent variable are termed univariate. Examples of univariate data sets include obtaining a single mass spectrum at one pixel within a MALDI image (the independent variable is m/z) or monitoring the intensity at one specific m/z across all pixels within an IMS experiment (the independent variable is location). Multivariate data sets contain more than one independent variable per measurement element. Examples of multivariate data include collecting entire mass
spectra at discrete points in time within separation methods (i.e. scan mode HPLC-MS where time and \( m/z \) are simultaneously two independent variables) or collecting entire mass spectra across several pixels within an IMS experiment (where location and \( m/z \) are simultaneously two independent variables). Principal component analysis (PCA) is an exploratory data analysis tool used to identify statistically significant sources of variation within multivariate data sets. Stated another way, PCA can be used to identify trends within data sets consisting of multiple independent spectra. A brief overview will be presented here as a guide for the data being presented, but readers are referred elsewhere for a full review [1,2].

The goal of PCA is to mathematically simplify the representation of a data set and can be best presented using an example. Consider a MALDI IMS data set compromised of individual mass spectra recorded for a set of pixels within an image. PCA identifies spectral features termed principal components (also called factors, eigenvectors, or loadings) that are common within pixels of a data set. Some of these principal components describe relevant information such as a peak at a particular \( m/z \) value while others only describe noise. Principal components are considered independent variables in PCA much the same way \( m/z \) values are considered independent variables in a recorded mass spectrum. In addition to loadings, PCA also identifies scores which are a set of scalar numbers that describe how important each spectral trend is to each data element. In terms of MALDI IMS data, scores represent how important each spectral trend is to each pixel. A large positive score value indicates that a spectral trend is strongly positively correlated with a particular pixel while a large
negative score value indicates that a spectral trend is strongly negatively correlated with a particular pixel. The entirety of a measured mass spectrum for a particular pixel can be reconstituted by summing the multiplicative products of each principal component with a corresponding score. A graphical view of this process is shown in Figure C.1.

Figure C.1: Example of PCA. Consider a spheroid slice analyzed by MALDI IMS. A mass spectrum taken from one pixel of a MALDI IMS data set (black, left) can be represented by a set of spectral features (colored, right) after performing PCA. PCA identifies several spectral features termed principal components (PCs) associated within all pixels of this particular data set (red, blue, green, and magenta graphs). A set of scores (numbers inset within each graph) describe how important each of these principal components are to the mass spectrum of this particular pixel. The original data can be reconstituted by multiplying each principal component loading by its score and summing the result.
Note that principal components must never be considered individual, pure analyte spectra that are somehow summed to give an overall recorded spectra; principal components are abstract representations that highlight trends present within the data and are not always directly interpretable [3,4].

The PCA algorithm successively calculates independent principal components in order of importance, with the first several principal components describing the prominent spectral features within the data and the latter principal components describing spectral minutia or noise. The principal components that describe relevant spectral information are retained and used for further data processing while those that describe noise are discarded with negligible impact as shown in Figure C.2. There are a variety of approaches for separating the principal components that describe relevant information from those that describe noise [2,5]. The Scree graph was used here [2,6]. In simple terms, this approach plots the relative amount of information (termed variance) described by each principal component. Principal components that are relevant should theoretically contain more variance than those that describe noise. The variance is calculated for each principal component, presented on a logarithmic scale, and an “elbow” in the graph is taken to be the maximum number of relevant principal components to retain.
Figure C.2: Noise Removal by PCA. The original spectrum (black trace) can be reconstructed nearly identically using only the first several principal components that describe relevant information (red trace). The latter principal components describe only noise (purple trace) and can be discarded.

A score plot is a graphical display often used with PCA. In this type of plot the magnitude of scores for each pixel along specified principal components are plotted (e.g. PC2 versus PC1). For MALDI IMS data, each point in this score plot represents the spectral content described by the plotted principal components for a particular pixel. Stated another way, since each principal component is a multivariate independent variable, the spectral content of a particular pixel is represented as a single point in an abstract multivariate coordinate system. Theoretically, the points associated with pixels that have similar mass spectral content should cluster together while points associated with pixels that have dissimilar spectral content should be segregated. This idea of transforming entire mass spectra into single points whose spatial distribution in a PCA score plot is associated with spectral similarity serves as the foundation for the clustering analysis (see below). Finally, much the same way the single points represent
spectra within a PCA score plot, any line drawn within a PCA score plot also represents an entire spectrum of information (since single points define a line). In this case, the axis lines representing principal components are the spectral trends identified by PCA. In sum, a PCA score plot of MALDI IMS data graphically plots the mass spectra of individual pixels according to how important/evident certain spectral trends (principal components) are exhibited within each of the pixels.

Mean-centering and variance scaling are two general data pre-processing steps that are often performed prior to PCA. Mean-centering is the process by which an average data spectrum is calculated and subtracted from each spectrum of the data set. Graphically, this process will center the data points within the PCA score plot about the origin. Without mean-centering, the first principal component calculated will simply be the average of the data set and will not describe variation in spectral content [7].

Mathematical scaling procedures are often used to ensure that all variables contribute equally during the PCA calculation. Intensities vary dramatically in mass spectra as a function of m/z. Higher intensity peaks typically have a numerically wider range of intensities across pixels as compared to lower intensity peaks [7]. Stated another way, the absolute standard deviation of peak amplitudes across all pixels for a common high intensity peak would be larger than the standard deviation of peak amplitudes across all pixels for a common low intensity peak. The m/z values associated with peaks that exhibit wide ranges in intensity describe more of the total variation of the entire MALDI IMS data set while m/z values associated with peaks that exhibit smaller ranges in intensity describe less of the total variation. Since PCA is a technique
that is used to identify sources of spectral variation in successive order of importance, performing PCA without scaling would cause high intensity peaks to dominate the calculated spectral trends/principal components at the expense of significant, yet lower intensity peaks.

A simple scaling method can be performed by dividing the intensity at a specified m/z for a specific pixel by the variance (square of standard deviation) of its intensity values across all pixels. Unfortunately, scaling often enhances noise. By treating all m/z values as equally important, m/z values that are associated solely with noise are given equal weight to the m/z values that describe relevant spectral features. Thus, variance scaling can sometimes degrade the quality of the data set being analyzed.

C.2 Cluster Analysis

The goal of cluster analysis is to identify and group similar observations present within a data set. Clustering is termed an unsupervised data analysis technique because the methodology is used without any a priori knowledge of the data set (e.g. the number of groups that exist, the proper groupings of elements within a data set, the structure/distribution of the data, etc.). Two clustering methods are discussed below: hierarchical clustering analysis (HCA) and \( k \)-means clustering.

In HCA, each point is initially treated as an individual cluster and inter-point distances are calculated for all data points. The two nearest points are merged together to be treated as a single entity, the inter-point distances across the data set are again calculated, and another two points are merged. This process continues in an iterative
fashion until all points are joined and only one nested cluster exists. The user then
estimates the number of clusters that exist based on a distance threshold where sets of
data elements must be a minimum distance apart to be considered separate clusters.
An excellent visual representation of this process is shown in Figures 4.10, 4.11, and
4.15 of Beebe [8].

There are several distance measures that can be used to join points together in
HCA. Depending on the spatial distribution of the data, sometimes these different
distance measures can generate alternative (although not necessarily incorrect)
groupings of the points within the data set [8]. Two common distance measures in HCA
include nearest-neighbor (termed single linkage) and furthest-neighbor (termed
complete linkage) [9]. Single linkage suffers from two disadvantages. First, grouping by
the nearest neighbor ignores the underlying distribution of the data set. Second, a
phenomenon termed chaining can occur whereby grouping errors arise because some
data points can be close enough to one another to allow grouping, yet can still be far
away from most of the elements of the group; graphically, HCA with nearest-neighbor
linkage can thus include points within a chain-link type distribution pattern away from a
center of mass. Analyses performed with complete linkage tend to bias towards creating
clusters with a spherical distribution [9].

\(k\)-means clustering is an iterative method that divides data elements into a pre-
defined number of groups (N) by minimizing the sum-of-squares spread of the cluster
about its center. This process is shown graphically in Figure C.3. Within the data set to
be clustered, N points are first randomly chosen as cluster centers (also called “seeds”)
and all other data elements are assigned into the group with the nearest cluster center. Next, the center of mass of the grouped elements (also called a centroid) is calculated and the process of assigning elements to the cluster with the nearest center is repeated. These two steps of calculating cluster centroids and assigning groupings are alternated iteratively until the result no longer changes (convergence is reached).

Figure C.3: \( k \)-means Clustering. A set of data was generated that contained two clusters of points with random noise added (A). The algorithm randomly chooses two cluster seeds to start the process (circled red and blue in A). All other points in the data set are assigned to the nearest cluster (B). Using these associations, new cluster centers are calculated, represented by the two Xs in B. The process of assigning points to clusters is repeated again as shown in C. Note how the cluster assignments changed. The calculation of cluster centers and point assignments is repeated until convergence is reached (D).
\textit{k}-means clustering suffers from two main disadvantages [10]. First, the assignment of data elements to clusters is highly dependent on the starting location of the cluster centers. To address this issue, the entire \textit{k}-means clustering process is repeated multiple times on the same data set using different starting seeds to ensure optimal clustering. Second, and more important, the number of clusters within the data set must be known and assigned before the clustering begins.

There are several methods that can be used to estimate the proper number of clusters without any a priori knowledge of a data set [9]. In the gap-statistic method, the spread (also termed dispersion) of points within groups of a clustered data set is compared against the spread of points within a data set containing random numbers [11]. As an example, assume a MALDI IMS data set ideally contains four separate groups of pixels that exhibit four unique mass spectral profiles. If fewer than four groups are used to cluster this “structured” data set, the dispersion of points within clusters will be large because not enough clusters exist to properly group the data points. In essence, the span of the too few clusters must increase to “reach out” farther to encompass all the data points. Conversely, if more than four groups are used, intact clusters are usually just split into smaller pieces, exhibiting a very small decrease to the within-cluster spread. For an unstructured (also called null) data set consisting of only random numbers that have no clusters, dispersion should gradually decrease at a constant rate (more or less) as the number of clusters used to group the data elements increase [11]. The “gap” between the dispersion of a data set consisting of random numbers and the
data set being tested is then used to estimate of the proper number of clusters for the test set.

C.3 Principal Component Analysis – Linear Discriminant Analysis

The field of discriminant analysis is traditionally used to classify unknown data elements into pre-defined groups based on data similarity. For example, assume MALDI IMS is used to monitor protein expression of four known cell types spatially localized within four distinct regions (e.g. north, south, east, and west) on the surface of an imaging plate. If each of these four distinct cell types contains a unique protein expression pattern, then (theoretically) four unique, characteristic mass spectral profiles should be generated. If these same four cell types are then randomly distributed across another imaging plate, discriminant analysis is an approach that can be used to associate each pixel of this new image with a specific cell type based on the similarity of their mass spectra with those unique, characteristic mass spectra previously recorded. Discriminant analysis is considered a supervised technique because information must be known by the user before classification can begin: the proper number of groups that exist within a data set and their corresponding group labels/identifiers (e.g. proliferating cells, necrotic cells, etc.).

Linear discriminant analysis (LDA) can best be described using an example. Consider the two spatially-segregated groups of points identified by k-means clustering shown in Figure C.3.1. The LDA algorithm calculates the equation of a line that best distinguishes between the two groups of data; this line (also called a classifier, linear
boundary, or linear discriminant) separates the graph into two distinct regions as shown in Figure C.4 [12]. The line calculated by LDA maximizes intergroup variance and minimizes intragroup variance, all with the lowest possible misclassification error. When any future observation of unknown association is plotted in this coordinate system, it would be classified as either one group or another according to its location on either side of the linear boundary.

Figure C.4: Calculation of a Linear Boundary by LDA. The data shown here is taken from Figure C.2.1(D). LDA calculates the equation of a boundary (black line) that segregates the coordinate system into two regions (orange for cluster 1 and green for cluster 2). Any future observations would be categorized into a specific group depending on its location on either side of the line.

As opposed to just a “line”, LDA more broadly refers to a linear combination of input variables (with exponents of one as opposed to a quadratic or cubic discriminant which would have exponents of two or three, respectively). In two dimensions, LDA
generates a line because a line is the linear combination of the two independent input variables $X$ and $Y$:

$$AX + BY = C$$  \hspace{1cm} (Eq. C.3.1)

where $A$, $B$, and $C$ are scalars ($A$ and $B$ are scalar coefficients) used to determine the direction of the line. This linear combination is not limited to only two measured variables. A plane would be created by LDA in a three-dimensional coordinate system because a plane is the linear combination of three input variables $X$, $Y$, and $Z$:

$$AX + BY + CZ = D$$  \hspace{1cm} (Eq. C.3.2)

While impossible to graph, LDA would continue to generate a linear-based boundary for data with larger numbers of input variables:

$$AX + BY + CZ + ...$$  \hspace{1cm} (Eq. C.3.3)

LDA will not provide a unique boundary for data sets that are over-determined, meaning that the data set could not contain more variables than data elements [13,14]. In terms of MALDI IMS data, LDA would normally require that the number of pixels exceed the number of $m/z$ values scanned within individual mass spectra. Unfortunately, this condition is rarely satisfied with MALDI IMS data sets as images would either need to contain thousands of pixels or the resolution within the mass spectra would need to decrease below the point of utility (e.g. binning mass spectra in increments of 250 $m/z$).

To review, PCA is considered a data reduction technique. Instead of constructing a mass spectrum using the recorded intensities of thousands of $m/z$ values, PCA can
construct an entire mass spectrum using far fewer independent variables (principal components) without changing the underlying content of the data as shown in Figure C.1.1. Because PCA dramatically reduces the number of independent variables, it can be used as a preliminary step prior to LDA, ensuring that data sets are not over-determined [13,14].

When PCA is used prior to LDA, the mathematical equation of the linear boundary determined by LDA would have the form

\[
(A \times PC1) + (B \times PC2) + (C \times PC3) + (D \times PC4) + ...
\]

(Eq. C.3.4)

where A, B, C, and D are scalar coefficients and PC1, PC2, PC3, and PC4 are the relevant principal components (note that equation C.3.4 would continue to include all relevant principal components of the data set). PCA offers another advantage prior to LDA in addition to the necessary data reduction step; noise within the mass spectra is discarded, improving the quality of the LDA boundary [13].

As stated earlier, a single point within a PCA score plot represents an entire data spectrum. Mathematically, each data spectrum is represented as a linear combination of principal components:

\[
\text{Spectrum} = (\text{ScorePC1} \times PC1) + (\text{ScorePC2} \times PC2)
\]

\[
+ (\text{ScorePC3} \times PC3) + (\text{ScorePC4} \times PC4)
\]

\[
+ ... \quad \quad \text{(Eq. C.3.5)}
\]
where specific score values generate the mass spectrum unique to a particular pixel within the MALDI IMS image. Much the same way that a recorded mass spectrum can be calculated using the linear combination of principal components and scores, a discriminant spectrum can be calculated using the linear combination of principal components and the scalar coefficients from the linear boundary (note the similarity in the mathematical formulas between equations C.3.3, C.3.4 and C.3.5). Because the linear boundary drawn maximally separates two groups of data (Figure C.3.1), the resulting discriminant spectrum will highlight spectral features that maximally differentiate the two groups of data [13].

C.4 Overall Data Workflow

A more detailed explanation of the overall statistical workflow process discussed in the main paper is presented here. First, the raw MALDI IMS data (recorded in the mzXML format) was loaded into the MATLAB programming environment. All mass spectra were originally recorded using an $m/z$ range of 4-20 kDa, but were trimmed to include only the relevant $m/z$ window. Spectra were then Gaussian-filtered to remove noise, a common default processing step available in many software packages that accompany commercial instrumentation15. Convolution with a Gaussian function can be used to make peaks more symmetrical and to reduce noise [16]. Finally, spectra were down-sampled to an $m/z$ resolution of ten $m/z$ units. Down-sampling allowed for significant improvements in data processing time and prevented errors associated with the limited memory capacity of the computer.
PCA was then performed on trimmed, filtered, down-sampled data to isolate relevant pixels. Mean-centering was used, but variance scaling was not used (see above) to ensure that noise present within the mass spectra would be of minimal influence in the PCA calculation. The spatial distribution of points within a score plot was used to isolate relevant pixels from the entire MALDI image (see Figure 4.2 of Chapter 4 and the Appendix B for a visual representation). This first iteration of PCA was used solely to remove irrelevant pixels and was not used to remove noise from individual mass spectra. All principal components (relevant and noise) were retained such that entire spectra could be processed subsequently.

Once the mass spectra of only the relevant pixels were isolated, the spectra were then further processed. Eventually a second iteration of PCA would be needed to isolate important spectral features within the relevant pixels of the image. Performing this second PCA step without variance scaling would bias the results towards high-intensity peaks (see above), yet variance scaling also would introduce noise which could dramatically degrade the quality of the results obtained by PCA. As a compromise, the mass spectra of the relevant pixels were logarithmically-transformed. This procedure reduced the absolute variation among intensities as a function of $m/z$ value such that PCA could identify spectral trends among high and low intensity peaks, yet did not magnify noise. While empirical, this approach has been used before successfully prior to PCA of MALDI IMS data [7]. A top-hat filter was then applied to the logarithmically-transformed spectra to remove the large baselines typically associated with MALDI matrix. Like Gaussian filtering, top-hat filtering is a routine step in the analysis of protein
mass spectra and is also available as a default processing options within commercial mass spectrometry software packages [15].

In a MALDI-TOF IMS experiment, subtle differences in sample height or sample roughness can cause subtle shifts in the recorded $m/z$ ratio of a protein so spectral alignment is essential prior to further processing [13]. Therefore, the next step used a simple interpolation-based aligning procedure on the logarithmically-transformed, filtered spectra to correct for slight $m/z$ shifts of peaks across the image [17]. Mass spectra were first “scooted” to align with a specific $m/z$ of a peak of low $m/z$ and then “stretched/shrank” to align around a second, higher $m/z$ peak. Afterwards, the logarithmically-transformed, filtered, aligned spectra were each normalized to unit area to correct for absolute differences in inter-pixel intensity. Finally, the processed spectra were mean-centered prior to a second round of PCA. This allowed for the easier interpretation of spectral trends, score plots, and PCA-LDA discriminant spectra (see below).

This second iteration of PCA was used to further remove spectral noise, reduce data dimensionality prior to LDA, and identify spatially-localized spectral trends present in the MALDI-IMS data set. As before, mean-centering was used and variance scaling was not. The Scree graph was used to determine the proper number of relevant principal components to retain [2,6].

After PCA was performed and the number of relevant principal components was calculated, the noise principal components were discarded. Next, the scores of the mass spectra along each relevant principal component were tabulated for each pixel. Finally,
the principal components and scores were visualized to identify spectral trends present within the data set (see Figure 4.3 of Chapter 4 and Appendix B).

Cluster analysis was then used to isolate groups of pixels according to similar spectral features present within the recorded mass spectra. In addition to its role in data reduction, PCA was performed prior to cluster analysis because the ability of PCA to remove noise improves the assignment of pixels to specific clusters [13]. HCA and \(k\)-means were both applied to identify cellular subpopulations within the MALDI-IMS image using the tabulated scores along each relevant principal component for each pixel as inputs. Stated another way, clustering was performed graphically within score plots (although all relevant principal components were used rather than just the first few that are traditionally visualized by researchers in score plots). Pixels were assigned into specific groups according to spectral similarity; these different groups were assumed to be cells within distinct microenvironments. The gap statistic was used for both HCA and \(k\)-means clustering to estimate the number of groups present in the dataset. The method of complete linkage was used as the distance measure for HCA clustering.

Once PCA was used to reduce data dimensionality and cluster analysis was used to segregate pixels into multiple groups containing unique cellular subpopulations, LDA was performed. LDA was used to generate the equations of linear boundaries between two user-specified cellular subpopulations. Once linear boundaries between groups were determined, discriminant spectra were calculated according to equation C.3.4. The interpretation of discriminant spectra is easier if the processed mass spectra are mean-centered prior to the second round of PCA [13]. If mean-centering is performed prior to
the PCA step, m/z ratios that exhibit positive intensities in the discriminant spectra are more strongly associated with one group, while m/z ratios that exhibit negative intensities are more associated with the other. Once calculated, the discriminant spectra can be used to identify specific m/z values selectively located within distinct microenvironments of the MALDI image.

C.5 References


