HETEROGENEOUS MOLECULAR SIGNATURES IN *STAPHYLOCOCCUS AUREUS* INFECTION ASSESSED BY MULTIMODAL IMAGING

By

Kavya Sharman

Dissertation
Submitted to the Faculty of the
Graduate School of Vanderbilt University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY
in
Chemical & Physical Biology
May 12, 2023
Nashville, TN

Approved:
D. Borden Lacy, Ph.D.
Brian O. Bachmann, Ph.D.
Tina M. Iverson, Ph.D.
Vito Quaranta, M.D.
Simon N. Vandekar, Ph.D.

ACKNOWLEDGMENTS

It would be impossible for me to try and thank each and every individual over the years who has encouraged, inspired, and motivated me in the pursuit of my education—but I shall do my best.

First and foremost, I would like to thank my advisor, Dr. Richard Caprioli, who saw a fire in me in our very first meeting and who has challenged me each and every day since to become a better scientist and educator. I would like to thank Dr. Jeffrey Spraggins for his continued mentorship, advice, and friendship. I would like to thank Dr. Raf Van de Plas, for not only challenging me intellectually when it came to data science and computation, but for taking the time to recognize the human aspects of being a student in science and challenging me to push through the marathon that is graduate school. I would also like to thank Dr. Heath Patterson for all his time, mentorship, and Python debugging. To my committee members, Dr. Borden Lacy, Dr. Brian Bachmann, Dr. Tina Iverson, Dr. Vito Quaranta, and Dr. Simon Vandekar—thank you for your mentorship over the years, which for many of you began from the time I was a Vanderbilt undergraduate student. You have helped me strive for the best while challenging me to be a better scientist, and I could not be more grateful.

My love for science began when I was a child, carrying around a Discovery Channel paperback book before I could properly read. For that and so much more, I would like to thank my parents, Sanjay Sharman and Dr. Jyotsna Sharman—I would not be here without your unending love, support, and inspiration. I would like to thank my sisters, Saumya and Ojasvini, who keep me smiling and encouraged throughout life. I would be nothing without you all.

I would like to thank all the teachers, educators, friends, and colleagues who have made a tremendous impact in my life. I especially have appreciated George Connell, Teena Smith, Beth Webber, Denise Euseppi, Connie Belk, Linda Johnson, Dr. Jon Ruehle, Dr. James Luba, Jill Cooper, Dr. Michelle Sulikowski, Professor Joseph Rando, Dr. Todd Giorgio, Dr. Michael Cooper, Dr. Gerardo Valadez, Dr. Ryan Ortega, Ariel Thames, Dr. Colin Walsh, Dr. Bruce Damon, Carolyn Berry, Dr. Elizabeth Bowman, Dr. Mark Frisse, Dr. Preston Campbell, Dr. Tim Hohmann, Dr. Ivelin Georgiev, Dr. Angela Kruse, Patty Mueller, Amanda Renick-Beech, Angie Pernell, Stryker Warren Jr., Dr. Charleson Bell, Dr. Ashley Brady, and Ann Gwin.

To all members of the Caprioli, Spraggins, and Van de Plas labs as well as those in the Mass Spectrometry Research Center and collaborating labs—thank you for your kindness and support. Thank you to all the graduate students, post-docs, and friends I have had the pleasure of knowing, both within Vanderbilt and the broader Nashville community. I would also be remiss if I did not acknowledge Vanderbilt University as a whole. I began my undergraduate journey in 2012 with a completely different vision of my future—the one I live today is not only better but more fulfilled and joyous than I could have imagined. I am so grateful for the community at Vanderbilt, especially Ms. Jackie at Last Drop, who was and still is a constant beacon of support and joy.

The research detailed in the pages below is a culmination of not only my scientific contributions over the past years, but the synergistic efforts of all the interactions I have had over the years, from brief conversations to years-long relationships. I am a systems biologist at heart, and I could not be more grateful for my system of family, friends, colleagues, and mentors—this dissertation exists because of you.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ABBREVIATIONS

µm: micron
AF: Autofluorescence
AQP1: Aquaporin 1
ATP: Adenosine Triphosphate
a.u.: Arbitrary Units
BALB: Bagg and Albino
CFU: Colony Forming Units
CO_2: Carbon Dioxide
CODEX: Co-Detection by Indexing
Cy5: Cyanine 5
DAN: 1,5-Diaminonaphthalene
DAPI: 4',6-Diamino-2-phenylindole
DBA: Dolichos Biflorus Agglutinin
ddH2O: Double-Distilled Water
DIA-MS: Data-Independent Acquisition Mass Spectrometry
DPI: Days Post-Infection
DsRed: Discosoma Red
EGFP: Enhanced Green Fluorescent Protein
H&E: Hematoxylin and Eosin
HMGB1: High Mobility Group Box 1
FDR: False Discovery Rate
FITC: Fluorescein isothiocyanate
FTU: Functional Tissue Units
HPLC: High-Performance Liquid Chromatography
IACUC: Institutional Animal Care and Use Committee
iBAQ: Intensity Based Absolute Quantification
ID: Identification
IF: Immunofluorescence
IMS: Imaging Mass Spectrometry
ITO: Indium Tin-Oxide
LC-MS/MS: Liquid Chromatography Tandem Mass Spectrometry
LED: Light Emitting Diode
m/z: Mass-to-Charge Ratio
LFQ: Label-Free Quantification
MALDI: Matrix-Assisted Laser Desorption Ionization
MicroLESA: Micro-Liquid Extraction Surface Analysis
mL: Milliliter
MS: Mass Spectrometry
MS/MS (MSn): Tandem Mass Spectrometry
MxIF: Multiplexed Immunofluorescence
NanoPOTS: Nanodroplet Processing in One Pot for Trace Samples
nm: Nanometer
NMF: Non-negative Matrix Factorization
OD: Optical Density
PANTHER: Protein Analysis Through Evolutionary Relationships
PAS: Periodic Acid-Schiff
PBS: Phosphate Buffered Saline
PC: Phosphatidylcholine
PCA: Principal Component Analysis
qTOF: Quadrupole-Time of Flight
ROI: Region of Interest
SAC: Staphylococcal Abscess Community
SM: Sphingomyelin
SVD: Singular Value Decomposition
sfGFP: Superfolder Green Fluorescent Protein
THF: Tetrahydrofuran
THP: Tamm-Horsfall Protein
timsTOF: Trapped Ion Mobility Time of Flight
TSA: Trypticase Soy Agar
TSB: Trypticase Soy Broth
UV: Ultraviolet
WCSS: Within Cluster Sum of Squares
WSI: Whole-Slide Image

CHAPTER 1 MULTIMODAL IMAGING AND *STAPHYLOCOCCUS AUREUS* INFECTION

OVERVIEW

Bioanalytical modalities such as imaging mass spectrometry, sequencing, and microscopy are commonly used for understanding biological processes of health and disease. Characteristics of each modality vary in terms of spatial resolution, molecular coverage, molecular specificity, and whether it is targeted or exploratory. For instance, stained microscopy provides high spatial resolution for visualizing cell and tissue-level structure but provides low molecular specificity. In contrast, matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry (IMS) provides label-free characterization of tens to thousands of chemical species within a single experiment; however, this comes at the cost of a coarser spatial resolution as compared to microscopy.

Combining information acquired from orthogonal imaging modalities allows for the creation of multimodal imaging datasets, providing an opportunity for untargeted data to be contextualized in terms of biomolecular pathways. These contextualized data can then be directed towards identifying molecular mechanisms of health and disease as well as diagnosis and prognosis. Methods that provide orthogonal information, such as mass spectrometry for chemical coverage and microscopy for spatial coverage, are best for generating multimodal imaging datasets.

Recent computational advances allow the combination of different imaging modalities through methods such as image registration and segmentation. However, extracting biologically meaningful information from multimodal imaging data remains a challenge largely due to the high dimensionality and chemical complexity of the data. There is an acute need for integrated computational methods so that we can link complementary imaging modalities and elucidate molecular findings using the context of one to mine the other. There is also a growing body of work that highlights the importance of shifting from traditional univariate approaches to more multivariate methods to provide systems biology levels of insight into the biological system at hand.

The ability to identify molecular species and map their spatial distributions in relation to known tissue substructures is a powerful way to track molecular differences associated with a specific disease and is a major goal of those using spatially targeted mass spectrometry. Analyzing mass spectrometry data collected from a tissue presents a number of challenges from high dimensionality to chemical complexity. The work summarized below aims to address and solve challenges associated with analyzing spatially targeted mass spectrometry data by developing new computational methods, data mining techniques, and visualization strategies.

The development of these new methods is of great importance to the field of imaging, leveraging powerful spatially aware methods alongside robust computational techniques to determine localized molecular changes and provide a systems biology-level summary of chemical changes in tissue. The application of these new computational methods enables a superior approach to analyzing multimodal imaging data than were previously possible.

Staphylococcus aureus soft tissue infections involve the formation of abscesses that cause changes in architecture within the host tissue microenvironment on a structural and molecular

level. By applying techniques developed within this body of work, the molecular heterogeneity of this bacterial infection was characterized and molecular drivers among staphylococcal abscesses and neighboring regions of infected tissue were elucidated.

INTRODUCTION

As technology advances, the amount of scientific information that can be produced from a given sample has increased enormously. This current era of big data offers incredible opportunities to understand health and disease with unprecedented nuance. However, it also presents unique computational challenges for the integration and visualization of datasets with high dimensionality and size.[1] It is essential to integrate multiple modalities of data to responsibly make conclusions about human health.

Bioanalytical modalities used to analyze tissue provide myriad types of data ranging from chemical information (comprising proteomics and lipidomics for instance) to spatial information (such as cellular and tissue neighborhoods obtained from stained microscopy). Each of these methods have trade-offs; methods that provide high spatial resolution are often limited in chemical specificity, whereas methods that provide deep chemical coverage have limited spatial information (Figure 1-1).

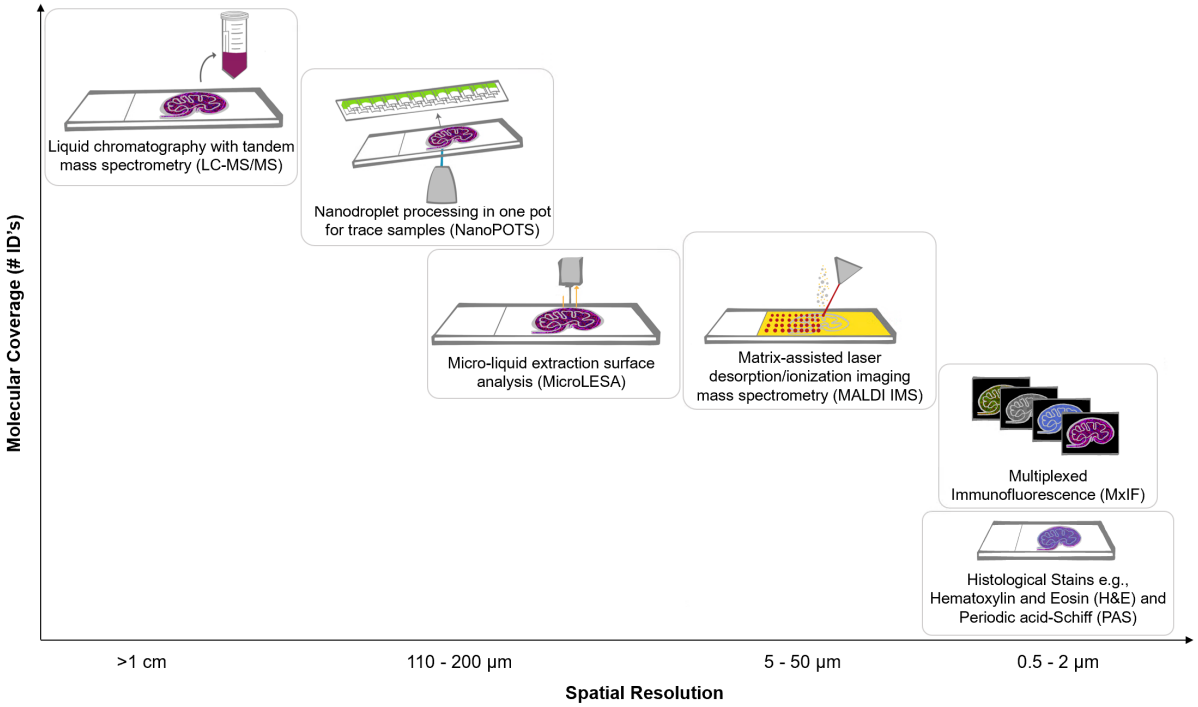

Figure 1-1: Spatially targeted biomolecular approaches, compared by molecular coverage (e.g., number of identifications) and spatial resolution (on a micron scale).

Of the methods detailed in this work, each provides a level of spatial resolution and chemical specificity (Table 1-1). Additional technologies along this spectrum are reviewed comprehensively in Prentice, et al. and Kruse, et al.[2,3] To capture both high spatial resolution and

high chemical specificity, the development of biocomputational methods to combine spatially localized molecular measurements from multiple imaging modalities is necessary.

	SPATIAL RESOLUTION	CHEMICAL SPECIFICITY	FOCUS
MALDI IMS	Coarse (e.g., > 10μm)	High	Exploratory
Multiplexed Immunofluorescence	Fine (e.g., < 1μm)	High	Targeted
Autofluorescence	Fine (e.g., < 1μm)	Low	Exploratory
Period-Acid Schiff	Fine (e.g., < 1μm)	Low	Targeted
Multi-modal Imaging	**Fine (e.g., < 1μm)**	**High**	**Exploratory + Targeted**

Table 1-1: Comparison of spatial resolution and chemical specificity for a subset of bioanalytical imaging modalities.

SPATIALLY LOCALIZED MOLECULAR MEASUREMENTS

Microscopy

Histological staining is a critical diagnostic tool and has been long been applied to characterize tissue features and cellular neighborhoods.[4,5] Staining protocols are often tailored to the specific tissue sample and experimental question,[6] although hematoxylin and eosin (H&E) and periodic acid-Schiff (PAS) are two broadly applicable staining approaches. Autofluorescence microscopy has recently been applied to characterize tissue architecture without disrupting the molecular composition of a sample.[7,8] Stained and autofluorescence microscopies offer high spatial resolution but lack chemical specificity.

Immunohistochemistry approaches apply an antibody to add specificity to tissue imaging.[9] Traditionally these approaches are limited in plexity due to the availability of primary antibody hosts and fluorescent reporters. Multiplexed immunofluorescence (MxIF) approaches such as Co-detection by indexing (CODEX) address this challenge by sequential quenching of fluorophores or the use of oligonucleotide barcodes.[10–12] These antibody-based techniques offer high spatial resolution but still lack the complex molecular information provided by mass spectrometry.

Spatially Targeted Proteomics

Within the field of proteomics, there have been advances to bridge the gap between spatial resolution and chemical specificity as well as chemical coverage. State of the art proteomics approaches usually involve liquid chromatography with tandem mass spectrometry (LC-MS/MS), which provides high proteomic coverage but does not maintain spatial information.[13,14] Typically, tissues are homogenized, samples are analyzed with LC-MS/MS, and large-scale proteomic differences among tissues are characterized. However, with the advent of spatially targeted approaches such as micro-liquid extraction surface analysis (microLESA),[15] we are now able to couple the high coverage of LC-MS/MS with spatial information.

Proteomics methods are subject to a trade-off between proteomic coverage (number of protein identifications) and spatial resolution (Figure 1-1). On one end of the spectrum, LC-

MS/MS of homogenized provides high proteomic coverage but lacks spatial information. Conversely, tissue imaging methods such as MALDI IMS provides proteomic data on a pixel-wise level at ~30-50 µm. Bridging that gap are hybrid technologies such as Nanodroplet Preparation in One pot for Trace Samples (nanoPOTS),[16–18] which can be combined with laser capture microdissection to extract samples for processing with LC-MS/MS, and microLESA,[15,19,20] which involves performing LC-MS/MS on selected ~100 µm regions of trypsin-digested tissue.

Analyzing data acquired from spatially targeted methods such as microLESA can be challenging due to the high dimensionality (few samples and thousands of proteins measured) as well as sparsity (missing values due to small tissue sample size), requiring custom data analytical pipelines.

Matrix-Assisted Laser Desorption/Ionization Imaging Mass Spectrometry

Matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry (IMS) allows for the untargeted detection of hundreds to thousands of molecular species from a tissue sample within a single experiment.[21,22] Although it has somewhat lower spatial resolution than microscopy, IMS enables the detection and identification of a wide range of biological species at increased molecular coverage with high spatial resolution and sensitivity using advanced data processing techniques. A typical MALDI IMS experiment is performed by first taking a tissue section and thaw-mounting it onto a glass slide (Figure 1-2). The tissue section is then homogenously coated with a crystallized matrix. An automated UV laser is then applied to the tissue in a raster pattern, ablating the tissue and allowing for acquisition of mass spectra at specific pixel locations corresponding to discrete x and y coordinates. Ion intensity heatmaps, commonly referred to as ion images, can then be generated for specific *m/z*, providing spatial localization and intensity information across the tissue.[21]

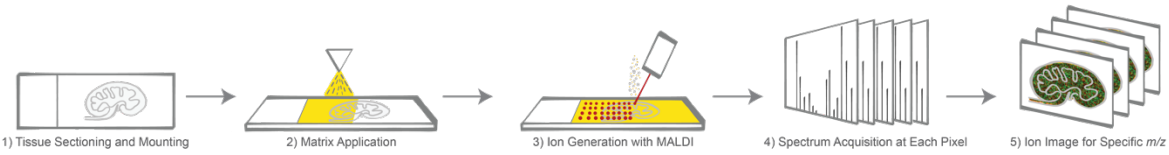

Figure 1-2: Schematic of a routine MALDI IMS experiment. 1) Tissue sections are cryosectioned and mounted onto a conductive glass slide. 2) MALDI matrix is applied uniformly across the slide. 3) An automated UV laser is applied in a raster pattern, generating a plume of ions that are analyzed by a mass spectrometer. 4) A full mass spectrum is generated at each pixel. 5) Ion intensity heatmaps, known as ion images, are generated for specific m/z *values.*

Although common methods and protocols for MALDI IMS have been developed, sample preparation protocols are often adapted for specific tissue types and analytes of interest to maximize detection of ions of interest.[23–27] Routine MALDI IMS experiments using commercial instruments are typically performed at a resolution of 10 µm, although laser focusing approaches and stage pitch reduction have enabled high spatial resolution as low as 1 µm.[28]

Several computational tools exist for the analysis of MALDI IMS data, but few are specifically designed to facilitate integration of MALDI IMS data with other modalities.[29,30]

BIOCOMPUTATIONAL METHODS FOR MULTIMODAL IMAGING

IMS data alone can provide spatially resolved molecular information. However, coupled with other types of imaging modalities such as microscopy, it allows for even higher spatial and molecular resolution. For example, IMS provides highly resolved chemical specificity, but relatively low spatial resolution. Microscopy can provide a much higher spatial resolution, but relatively low chemical specificity. As such, combining an IMS dataset with a microscopy image, each obtained on serial sections of tissue can offer much higher resolution chemical and spatial resolution.

Recent advances to couple these IMS and microscopy include multimodal image registration.[8] Landmark registration has been used to co-register MALDI IMS data with stained and autofluorescence microscopy. The resultant registered image offers improved spatial resolution for MALDI IMS data and allows for more granular comparison of tissue substructures.[8] Another example of multimodal imaging can be seen with image fusion of IMS and microscopy images.[31] This computationally driven process integrates IMS and microscopy, resulting in an image that is rich in chemical and spatial resolution using a statistical regression approach. Multivariate regression is applied to microscopy measurements to predict ion distributions, increasing ion image resolution by an order of magnitude.

Image registration and fusion are two methods that can be performed using IMS and other imaging modalities that offer higher chemical and/or spatial resolution. Ongoing efforts at the interface of hardware and software development seek to further improve chemical specificity, chemical coverage, and spatial resolution by improving IMS technology as well as incorporating other imaging modalities such as multiplexed immunofluorescence, transcriptomics, and elemental imaging.

Computational challenges for multimodal imaging

Several challenges exist for the analysis and integration of multiple data types. These include data dimensionality and structure as well as data visualization.

IMS data is highly dimensional and large in terms of data size, often requiring a form of data compression to decrease the computational load.[29,30,32] Methods to compress IMS data include binning mass spectra for each pixel or compressing them based on regions of interest (ROI) and generating average spectra. One compression method based on ROIs is image segmentation, which involves subdividing tissue regions with homogenous spectral profiles and identifying co-localized *m/z* values.[29,30] Another data compression approach uses unsupervised machine learning to reduce dimensionality and extract features (specific *m/z* values) for downstream statistical analysis.[32] One common unsupervised approach is principal component analysis (PCA), which involves combining individual variables (specific *m/z* values) into linear combinations that exhibit similar behavior. These groups are known as "components" and represent patterns in the data with far fewer dimensions. Another example of an unsupervised approach is non-negative matrix factorization (NMF), which works similarly as PCA, with the exception that all calculations are performed in the positive domain, making it especially useful for IMS observations where there are no negative values in the dataset and the resultant components can be represented as average spectra. In addition to these types of dimensionality reduction techniques, other unsupervised machine learning approaches such as hierarchical clustering and *k*-means clustering can also be used to determine potential groupings of the data based on spectral similarity.[32]

In addition to data dimensionality and structure, visualization of IMS data remains a challenge. Large and multivariate datasets require extensive computational power and visualizing multiple imaging modalities in an integrated manner further necessitates computational bandwidth. Advanced machine learning methods have been developed to mathematically integrate IMS data with microscopy into a combined form; these include data-driven image fusion[31] and interpretable machine-learning based marker discovery.[33] However, these integrated images are not as functional for molecular imaging studies where human interpretation by domain experts is the goal. Although image viewers such as QuPath[34] and Napari[35] support multiple imaging modalities where viewers can import multimodal imaging data and selectively view each side-by-side or overlaid with varying degrees of transparency, there remains a need for novel visualization strategies that allow for spatial mapping of different source modalities into the same coordinate systems while still allowing the content of the original modalities to be viewed and considered separately.

CASE STUDY: *STAPHYLOCOCCUS AUREUS* ABSCESS FORMATION AND DEVELOPMENT

Staphylococcus aureus is a gram-positive bacterium which presents a severe public health concern, responsible for over 20,000 deaths[36] and costing between $3.2 billion to $4.2 billion[37] annually in the US alone. A hallmark of *S. aureus* infection and disease progression is the formation of abscesses.[38] These abscesses begin as bacterial colonies and quickly progress into intricate three-dimensional structures that cause changes in architecture within the host tissue microenvironment on a cellular and molecular level.[39,40] These abscesses, once thought to be static lesions, are now understood to be dynamic environments consisting of a staphylococcal microbiology at the center of the abscess with defined layers consisting of necrotic host tissue, host immune cells, and microbial cells which release factors that support disease progression.[19,39,40]

The geometry of the abscess changes depending on the region of infection, making each infection a unique biological system. Further, the three-dimensional nature of these infections presents more complexity because the differential interaction between the bacteria and their host environment results in depth gradients of oxygen and nutrients necessary for their survival as well as gradients in host and pathogen proteins involved in inflammation.[39,41] Previous work has suggested that these abscesses are simply regions of dead neutrophils and stagnant bacteria; however, recent work has shown that these abscesses are in fact comprised of active bacteria with spatially oriented gradients of living and dead immune cells.[42] These gradients indicate that there is a complex microenvironment within the host-pathogen interface, where presumably there is a complicated interplay between the host proteins and pathogen proteins. For instance, it has been speculated that intracellular *S. aureus* could be a reservoir for antibiotic resistant bacteria.[42] Combined, these factors make spatial characterization of the bacterial host-pathogen interface a major challenge.

SUMMARY AND RESEARCH OBJECTIVES

In order to improve our current biocomputational approaches to integrate multimodal molecular imaging data and further understand staphylococcal abscess development, a series of technological and biological advancements were made using three independent approaches.

First, an automated unsupervised method for analyzing high-dimensional spatially targeted proteomic data utilizing PCA followed by *k*-means clustering was developed. The technical contribution was to create a method to rapidly filter complex proteomics data from a microLESA experiment and determine the most relevant species from hundreds to thousands of measured proteins in the form of ranked protein lists and pathway enrichments, thereby providing a systems-level view into complex molecular biological processes. From a biological standpoint, this method was used to identify key metabolic and cytoskeletal reorganization processes involved in infection as well as proteins involved in calcium-dependent, metabolite interconversion, and cytoskeletal processes that were enriched in sites of infection, especially at the ten days post-infection timepoint.

Second, CODEX immunofluorescence and MALDI IMS data were integrated and segmentation methods for microscopy were evaluated involving watershed and intensity-based thresholding; upon evaluation, it was determined that a more multivariate approach was needed and so a *k*-means clustering method for CODEX immunofluorescence data was developed. From a biological perspective, cell types that were not labeled by antibodies were discovered and spectra for these cell types were extracted and molecular heterogeneity within a staphylococcal abscess was observed; lipids co-localizing to specific abscess rich and non-abscessed regions were also identified.

Finally, to address the challenge of multi-modal microscopy and IMS data visualization, a contour mapping strategy was developed to overlay whole-slide images comprising IMS and PAS. Biologically, this provided insights into the directionality and morphology of abscess development using IMS signals in the form of contour maps.

CHAPTER 2 A RAPID MULTIVARIATE ANALYSIS APPROACH TO EXPLORE DIFFERENTIAL SPATIAL PROTEIN PROFILES IN TISSUE

OVERVIEW

Spatial proteomics is a method that analyzes the protein content of specific cell types and functional regions within tissue. Although spatial information is often key to understanding biological processes, interpreting region-specific protein profiles can be challenging due to the high dimensionality of the proteomic data acquired. Herein, we developed a multivariate analysis approach to rapidly explore differential protein profiles acquired from distinct tissue regions and applied it to analyze a spatially targeted proteomics dataset collected from *Staphylococcus aureus*-infected murine kidney at two timepoints (4- and 10- days post infection). This approach consists of applying a principal component analysis (PCA) for dimensionality reduction of protein profiles measured using micro-liquid extraction surface analysis (microLESA) mass spectrometry. Following PCA, *k*-means clustering was applied onto the PCA-processed data, thereby grouping samples by chemical similarity in an unsupervised manner. Cluster center interpretation revealed a subset of proteins that differentiate between spatial regions of infection over two time points. A gene ontology analysis of these proteins revealed that these proteins are involved in metabolomic pathways, calcium-dependent processes, and cytoskeletal organization. We also discovered that Annexins 2, 3, and 5 were increased in areas of infection and speculated that while Annexin 2 may be facilitating staphylococcal anchoring in tissue, Annexin 3 and Annexin 5 may be conferring various degree of host protection during infection. In summary, applying this multivariate analysis pipeline to an infectious disease case study highlighted differential protein changes across regions of infection over time, highlighting the dynamic nature of the host-pathogen interface.

INTRODUCTION

The proteomics field has developed an extensive set of methods to separate, purify, identify, and quantify proteins.[43–48] For instance, liquid chromatography with tandem mass spectrometry (LC-MS/MS) was developed to provide deep proteomic coverage on the order of tens to thousands of proteins; however, this coverage is possible due to tissue homogenization, which removes spatial context. [13,14] As such, LC-MS/MS is powerful for identifying proteins and their post-translational modifications. Within the biomedical space, this method has been applied to research diseases such as cancer[49–51], diabetes[52–55], and heart disease[56–59]. Proteomics can also be applied in a spatially targeted way, for instance, using matrix-assisted laser desorption/ionization imaging mass spectrometry (MALDI IMS),[21,22,60–62] which can be used to acquire protein measurements for hundreds of species simultaneously at a relatively high spatial resolution of anywhere from 10μm to 50μm.[63,64] As a result, MALDI IMS can provide an unparalleled combination of high plexity and high spatial resolution for molecular imaging; however, since the sample acquisition is over a smaller section of tissue as compared to a larger homogenized tissue section, the overall protein coverage and confidence in identifications is lower than LC-MS/MS.[65,66]

In between these two extremes in terms of spatial and protein coverage are hybrid technologies that leverage histology-directed spatial acquisition with LC-MS/MS. As a result, these hybrid methods can provide deeper molecular coverage than other spatial analyses from sampled regions. One such method is Nanodroplet Processing in One Pot for Trace Samples (nanoPOTS).[16–18] NanoPOTS involves the use of laser capture microdissection for spatially targeted sample acquisition followed by LC-MS/MS analysis for protein identification and quantitation.[67,68] This method routinely provides an average of 2,000 protein identifications at a 100μ spatial resolution.[18] Micro-Liquid Extraction Surface Analysis (microLESA) is another hybrid technology which involves histology-driven selection of regions of tissue. Image-guided robotic spotters are used to deposit picoliters of a proteolytic enzyme solution on regions of interest that are about ~100μm in diameter (for reference, a traditional LESA experiment is usually performed on regions that are 1-2mm[15,69–71]). After an incubation period, proteolytic peptides are extracted and analyzed using LC-MS/MS.

As with any method, spatially targeted LC-MS/MS methods have their unique set of challenges. After LC-MS/MS acquisition, the data are processed using software such as MaxQuant,[72,73] which cross-references the mass spectra with reference spectral databases, performs peptide and protein quantification, and finally generates protein identifications for each sample. Since these methods target smaller tissue regions, the amount of tissue acquired for analysis is reduced, thereby resulting in an overall decrease in the total number of proteins detected as compared to standard bulk proteomics results. This generates missing values within the dataset. Other reasons for missing values can be due to protein concentrations being below the limit of detection, protein measurements being filtered based on user-defined criteria, or proteins being missing randomly due to technical issues or a borderline signal-to-noise ratio. Furthermore, there is often molecular heterogeneity present in the samples due to biological variation based on sample location, which leads to differences in the total number of proteins identified per sample as well as representation of specific protein families. Within this particular study, protein coverage ranged from 31% - 77% among samples.[15,74]

Current methods for analyzing proteomics data are often univariate in nature, focusing on individual proteins and assessing their differential expression among samples and between disease states.[75–83] Although powerful in their own respect, these methods tend to be less applicable for capturing systems-level trends or panels of molecules working in unison, thereby limiting their effectiveness at retrieving the most information from originally complex multivariate data. This aspect coupled with the missing values prevalent in spatially targeted data make these data unamenable to one-on-one protein comparisons without pre-processing steps such as imputation.[84–86] Supervised methods can be applied for protein studies; however, these are largely applied for categorizing samples of tissue into categories such as diseased and non-diseased or to differentiate among tissue regions.[87–89] In comparison, the advantage of an unsupervised approach is that the data are allowed to separate into underlying trends, some of which will be non-biological and others biological, as opposed to a supervised approach where the analysis focused on recognizing specific pre-determined categories.

In this work, we address the challenges and nuances of spatial proteomic data analysis by describing the development of a rapid automated unsupervised multivariate method using PCA and *k*-means clustering to discover molecular differentiators within a publicly available microLESA data set investigating *Staphylococcus aureus* infection in a murine kidney on a spatial scale and over two timepoints.[74] This staphylococcal model was selected to provide insight into

the profound protein changes within tissue that contain bacterial abscesses, while maintaining broad multivariate protein coverage and avoiding prior focus on specific tissue classes of protein species.

RESULTS & DISCUSSION

S. aureus is a Gram-positive bacterium known to cause skin and soft tissue infections.[40] A hallmark of staphylococcal infection is the formation abscesses within soft tissue, the development of which is often accompanied by changes in host architecture as well as changes in host cellular and molecular composition.[38–40,74] Elucidating the formation of these structures, which include tracking the molecular changes across different regions of, and in proximity to, the abscess is critical for understanding how *S. aureus* interacts with the host immune system to infiltrate and proliferate as abscesses.

In the originally published experiment,[74] mice were infected with fluorescently tagged strains of *S. aureus* and their kidneys were excised for analysis at 4 or 10 DPI (Figure 2-1). Three regions were selected for extraction[74]: the staphylococcal abscess community (SAC), the non-abscessed cortex, and the interface between the abscess and the surrounding non-abscessed cortex (Figure 2-2A). 42 samples were collected altogether, with 20 collected at 4 DPI (5 from the interface, 7 from non-abscessed cortex, and 8 from the SAC) and 22 at 10DPI (6 from the interface, 7 from non-abscessed cortex, and 9 from the SAC). There were 3 biological replicates for each DPI category (6 mice total, 3 mice at 4DPI and 3 mice at 10DPI) and multiple ROIs from each of the 3 regions were sampled for analysis.

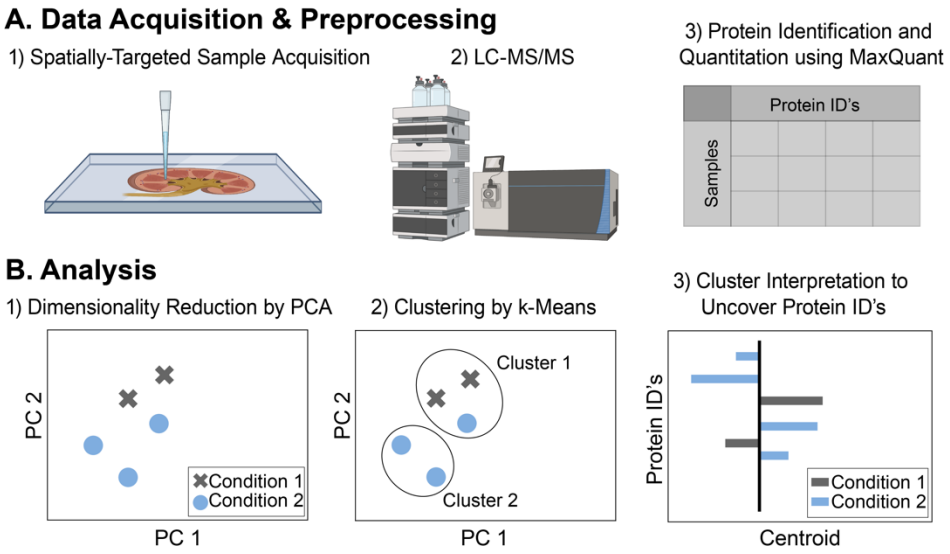

Figure 2-1: Pipeline for Spatially Targeted Proteomics Data Acquisition and Analysis. A) Protein data were acquired from tissue samples using spatially targeted sample acquisition and then peptides were analyzed using LC-MS/MS. Data preprocessing involved protein identification and quantitation using MaxQuant software. B) PCA was applied for dimensionality reduction and grouping of correlated and anticorrelated proteins among regions and timepoints. The PCA-processed data were clustered by k-means, and cluster centers examined for protein identifications.

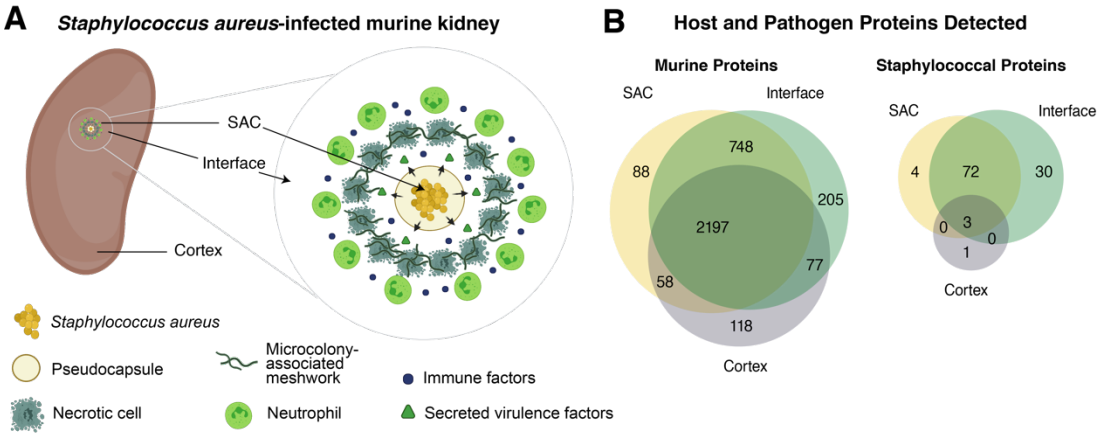

Figure 2-2: S. aureus-infected murine kidney. A) Graphical depiction of the host-pathogen interface of S. aureus infection within a murine kidney. SAC: staphylococcal abscess community (SAC). B) Summary of the total number of host and pathogen proteins detected.

Protein Identification and Quantification

Approximately 1,500 proteins were detected from each sample. Proteins unidentified in one sample but detected in another sample or that were below the limit of detection were not reported, thereby generating a missing value in the LFQ table. The percent of missing values in each sample ranged from 31% to 77% (Table 2-1).

Table 2-1: Missing values per sample

SAMPLE	PERCENT MISSING VALUES
4DPI_interface_2_4	52.71
4DPI_interface_2_2	47.24
4DPI_interface_2_1	52.32
4DPI_interface_1_2	77.21
4DPI_interface_1_1	59.70
4DPI_cortex_3_3	56.44
4DPI_cortex_3_2	60.77
4DPI_cortex_2_3	57.73
4DPI_cortex_2_2	53.29
4DPI_cortex_2_1	58.70
4DPI_cortex_1_2	65.17
4DPI_cortex_1_1	50.33
4DPI_SAC_3_3	46.19
4DPI_SAC_3_2	48.78
4DPI_SAC_2_3	71.71
4DPI_SAC_2_2	72.98
4DPI_SAC_2_1	36.96
4DPI_SAC_1_3	45.91
4DPI_SAC_1_2	59.45
4DPI_SAC_1_1	48.76
10DPI_interface_5_3	31.02
10DPI_interface_4_3	38.92
10DPI_interface_4_2	36.35
10DPI_interface_4_1	46.27
10DPI_interface_3_2	36.46
10DPI_interface_3_1	37.27

10DPI_cortex_4_4	64.61
10DPI_cortex_4_3	72.21
10DPI_cortex_3_3	63.84
10DPI_cortex_3_2	58.90
10DPI_cortex_2_3	95.41
10DPI_cortex_2_2	92.87
10DPI_cortex_2_1	89.42
10DPI_SAC_4_3	41.74
10DPI_SAC_4_2	47.90
10DPI_SAC_4_1	51.77
10DPI_SAC_3_3	42.32
10DPI_SAC_3_2	52.82
10DPI_SAC_3_1	41.05
10DPI_SAC_2_3	63.40
10DPI_SAC_2_2	55.33
10DPI_SAC_2_1	66.22

Outliers were assessed by calculating z-scores for each sample based on the number of protein groups identified in each and excluding samples with a z-score > |2|. Three samples were excluded and a table of 39 remaining samples was generated containing protein group versus LFQ intensity. Missing protein values can be handled in several ways, including imputing them based on a pre-defined model or removing proteins that were not detected across all samples. For our primary analysis, we opted for the latter; all proteins with missing values in one or more samples were excluded, leaving a total of 287 protein groups. All 287 protein groups were identified as murine using the MaxQuant database search. Of note, although in this case study MaxQuant LFQ was used as the input data, other value types such as iBAQ intensities or raw ion intensities can also be provided as input data for this PCA + *k*-means workflow, without it requiring substantial changes. The choice of which input type to supply depends on what is most appropriate for the data set and analysis at hand.

Once the data were pre-processed, we sought to develop an unsupervised multivariate method that would allow us to capture the unique proteomic signature from each of the distinct ROIs, but without focusing on a specific protein species and instead providing broad coverage across a panel of proteins. The entire 287-protein-group data set was used (4DPI (n=20) and 10DPI (n=19)) and region (SAC (n=17), interface (n=11), or cortex (n=11)). The data were not pooled by technical or biological replicate in order to provide as many measurements as possible for the unsupervised learning and to avoid an "averaging out" of information, which has the potential of underpowering the analysis, especially given the low number of samples already (n=39).

Principal Component Analysis Followed by *k*-Means Clustering

PCA was selected to address the "curse of dimensionality," which broadly summarizes the myriad challenges in analyzing and identifying patterns in high-dimensional data. PCA works by grouping correlated and anticorrelated features into a series of orthogonal components. As a result, the data are transformed from a high-dimensional space into a lower-dimensional space with minimal loss of information.

For this analysis, PCA with a randomized solver[90] was applied to reduce the dimensionality of the data and group correlated and anticorrelated proteins based on the protein LFQ intensity values. This resulted in a reduction of the overall dataset, from a matrix of dimensions [39 x 287] to that of [39 x 39]. All components were retained to avoid the loss of information. The first and second principal components accounted for 81.35% and 10.61% of the explained variance, respectively, and together, these components separated the data by region and timepoint (Figure

2-3A). We further labeled the data by regions and timepoints to explore variation present among these subsets. (Figure 2-3A, C). Samples collected from the (uninfected) cortex cluster seem to separate away from those collected from the interface and SAC, suggesting similarity between the interface and SAC proteomics, which is expected since both contain regions of infection (Figure 2-3A). There is also a degree of protein similarity among the biological replicates because samples within the PCA seem to cluster similarly based on biological replicate (Figure 2-3B).

Studying the results of the PCA as a function of time point reveals a clear distinction between samples collected 4 and 10 DPI (Figure 2-3C). This suggests that infection time is a key differentiator among the protein patterns in both interface and SAC regions. Some interface samples that were collected 10DPI overlap closely with SAC samples collected 10DPI, suggesting that, after 10DPI, the interface proteome could potentially start resembling that of the SAC. This observation is indicative of interface heterogeneity and a differential impact of infection among regions of tissue surrounding bacterial abscesses. It may also imply spatial expansion of the immune response and expanding tissue damage as result of the progressing infection, but this finding would require subsequent follow-up study and validation. Conversely, samples acquired from the cortex where there was no infection visibly present do not show a separation between 4 and 10DPI.

Although PCA provides groupings of samples based on the protein (LFQ) content, a secondary step is required to identify and interpret protein patterns. The hypothesis was that an automated unsupervised clustering method would provide additional insight into the spatial patterns of *S. aureus* infection over early and late time points. Unsupervised clustering is commonly applied to high-dimensional data as it involves grouping similar samples together based on variation among measured features. Within a protein data set, samples that contain similar protein expressions patterns are grouped together and the underlying variation among the groupings can potentially represent relevant information. There are many methods for clustering,[32,91,92] but *k*-means clustering[93–95] with a Euclidean distance metric was chosen because the cluster centroids, representing the average protein pattern for each group can provide rapid and straightforward protein-level insight.

For this non-imputed data set, *k*-means clustering was applied to the PCA-transformed data. Silhouette scores[96] were used as a performance metric and a *k* of 4 was selected. Each sample was subsequently assigned membership to 1 of 4 clusters descriptors of samples in each cluster were added to aid in interpretation (Figure 2-3D).

Studying the results of the clustering, samples from the non-abscessed cortex are grouped into cluster 3 and include samples from both 4DPI and 10DPI. However, there was varied cluster membership for samples extracted from areas of infection. cluster 4 is composed of samples acquired from the cortex and interface 4DPI and 10DPI, which suggests a similarity between the proteome of the interface early in the infection and the cortex as opposed to the proteome of other interface samples or the SAC. Cluster 2 consists of interface and SAC samples, both collected at 4DPI and 10DPI. Conversely, cluster 1 only contains samples from interface and SAC at 10DPI. The clustering patterns of interface and SAC samples lead to two interesting observations. First, the SAC samples separating into two different clusters is consistent with findings of previously observed abscess heterogeneity[39,74] and indicates that there can still be changes in abscesses that are seemingly fully formed. Second, that abscess formation and mounting immune response may take up to 10DPI to manifest in proteomic changes within the interface even ugh abscesses can be seen after 4DPI. This distinction in the protein content between the early and late interface is

notably evident as interface samples are seen clustering with cortex and SAC samples at varying timepoints. In summary, these observations from the *k*-means reveal patterns of staphylococcal infection progression and heterogeneity in the proteome among specific regions of infection.

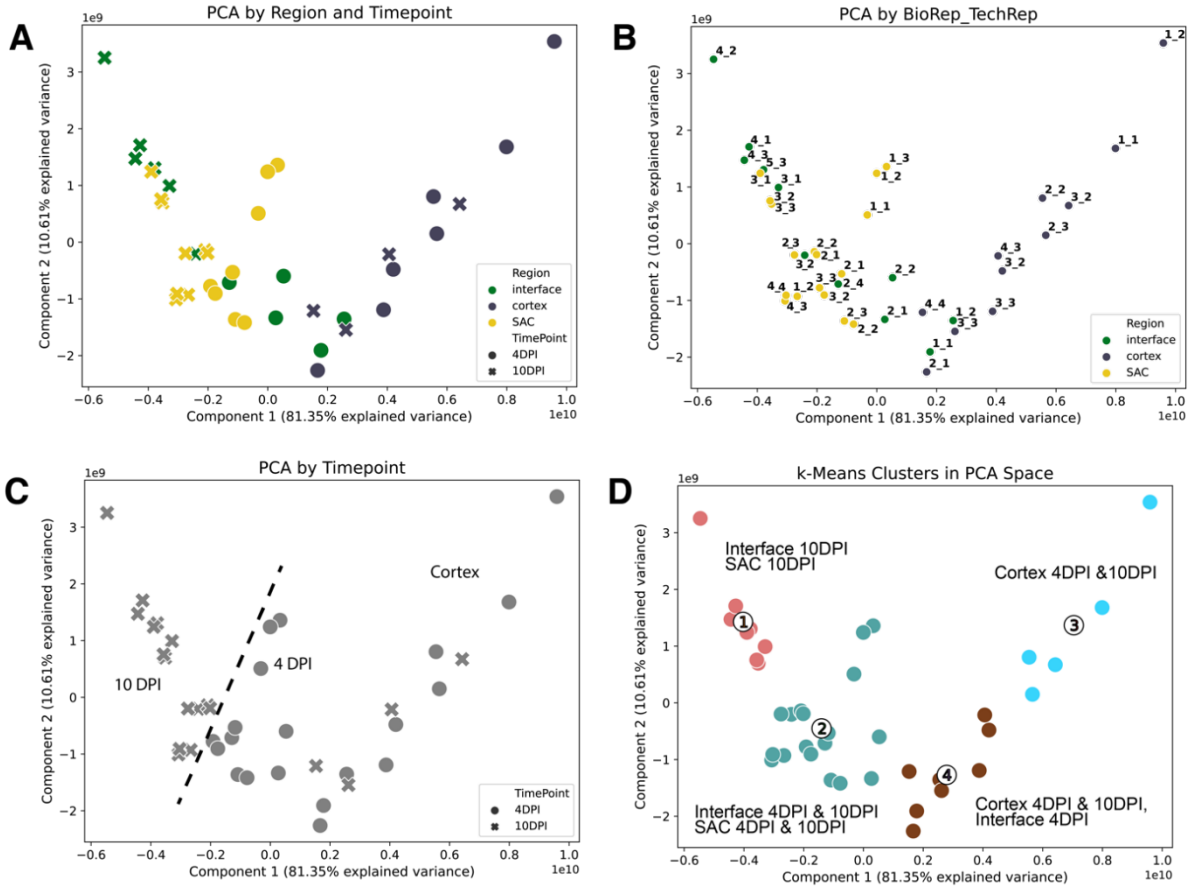

Figure 2-3: Principal Component Analysis (PCA) and k-means clustering results of proteins of an S. aureus infected murine kidney. A) PCA was performed on protein LFQ intensity values acquired from 3 regions and 2 timepoints. This unsupervised approach separates the SAC and interface (left) from the cortex samples with no visible infection (right). B) Samples also seem to cluster based on biological replicate within the PCA space. C) There is a separation among the samples 4- and 10-days post infection within samples acquired from region of infection; this separation is not seen from samples acquired from the cortex where there was no visible infection. D) k-means clustering was used to cluster the samples after PCA. k = 4 was determined using silhouette scores as a metric. To aid in interpretation, clusters are labeled by the regions and timepoints from which samples were collected.

Cluster Interpretation

The multivariate analysis presented provides broad proteomic insight based on underlying proteomic (LFQ) variations and, within this model of infectious disease, offers a high-level understanding of protein differences between two time points and three regions within a soft-tissue *S. aureus* infection. In order to identify a subset of relevant proteins from the total 287 proteins measured, the average centroids of each cluster were analyzed. The PCA followed by *k*-means approach allows for the automatic ranking of proteins that significantly contribute to the clustering

models such that proteins with high absolute centroid values contribute more as differentiators within the clustering model as compared to those with low absolute centroid values, thereby potentially signaling biological relevance. The average centroids for each of the four clusters were extracted, with 287 proteins or protein groups as observations and absolute cluster centroid values as variables. Each centroid includes all proteins and proteins with high absolute values are more relevant to a given cluster than those with low values.

Of the 287 proteins analyzed in the non-imputed dataset, the top 10% with highest absolute centroid values were labeled for interpretation (Figure 2-4). Alpha globin 1, cytoplasmic actin, and beta-globin have the highest absolute centroid values and distinguish clusters comprising samples from infected regions (interface and SAC) as opposed to those from the cortex. Also among these top 10% of proteins are mitochondrial ATP synthase subunits alpha & beta and pyruvate kinase, which are involved in ATP synthesis, as well as proteins involved in maintaining cell structure and facilitating tissue repair/remodeling such as myosin-9, cytoplasmic actin, filamin, and fibrinogen.

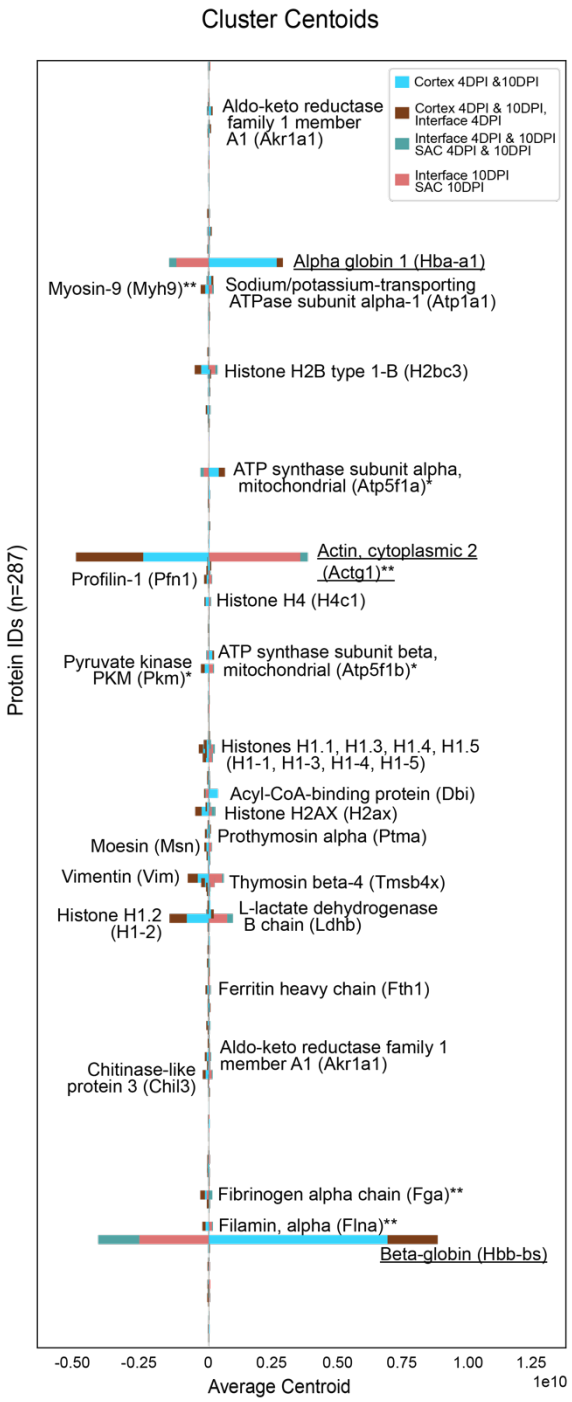

*Figure 2-4: Molecular differentiators among regions of S. aureus-infected kidney. A) All four cluster centers are overlaid, with 10% of proteins (n=29) with highest absolute values labeled. Underlined are the three proteins with overall highest absolute centroid values. * = proteins involved in ATP synthesis. ** = proteins involved in maintaining cell structure and facilitating tissue repair/remodeling.*

Analysis on Imputed Data Set

Given the low number of samples (n=39) covering two time points and three regions, eliminating proteins with missing values in one or more samples resulted in eliminating a substantial amount of measured protein data. We re-analyzed the data set, this time with all the

protein data and the incorporation of an imputation approach to handle the missing values. Imputation has been systematically evaluated and successfully implemented for mass spectrometry datasets and studies.[84,90,97] A recent study demonstrated that techniques such as local least-squares, random forest, and Bayesian PCA missing value estimation work well for label-free data-independent acquisition mass spectrometry (DIA-MS) datasets.[85] However, this study among others revealed that for an imputation to be biologically relevant, it must model actual observed phenomena. Within the case of microLESA data where proteins are sampled from small, biologically heterogeneous regions of tissue, the probability that a protein was not detected when its value is missing is higher than if we were to impute a value based on an imputation method. Therefore, for this case study, we chose a simple model with the assumption that if a protein was not measured, it was below the limit of detection or not present in the sample. Therefore, missing values were zero-filled. This type of imputation is a common approach for handling missing value, as opposed to our primary approach, where columns with missing values were removed such that only globally present proteins are used for the analysis.[84,85,97]

This zero-filled data set comprised 3,613 proteins in total. The PCA and *k*-means clustering results remained largely the same (Figure 2-5) with notable exceptions: within the PCA, components 1 and 2 now respectively represent 77.26% and 10.3% of the data as compared to the 81.35% and 10.61% previously, Figure 2-5A-C), silhouette score analysis revealed a *k* of 5 to be optimal for this larger richer dataset with imputed values, (Figure 2-5D), and resultant cluster membership of samples. Furthermore, within the *k*-means clustering results, there is a new cluster intermediately situated between the cortex and interface/SAC 4 and 10DPI clusters. This new cluster consists of samples from the cortex 4 and 10DPI as well as interface 4DPI. Samples originally organized into a single cluster comprising cortex 4 and 10DPI split into two clusters, with two samples acquired from the cortex at 4DPI comprising one cluster and six samples acquired from the cortex at 4 and 10DPI comprising a second cluster. This change from the original clustering output indicates that with the inclusion of all proteins and zero-filling those with missing values, we can observe more perceived separation among the samples. The cluster with interface and SAC samples 10DPI remains unchanged.

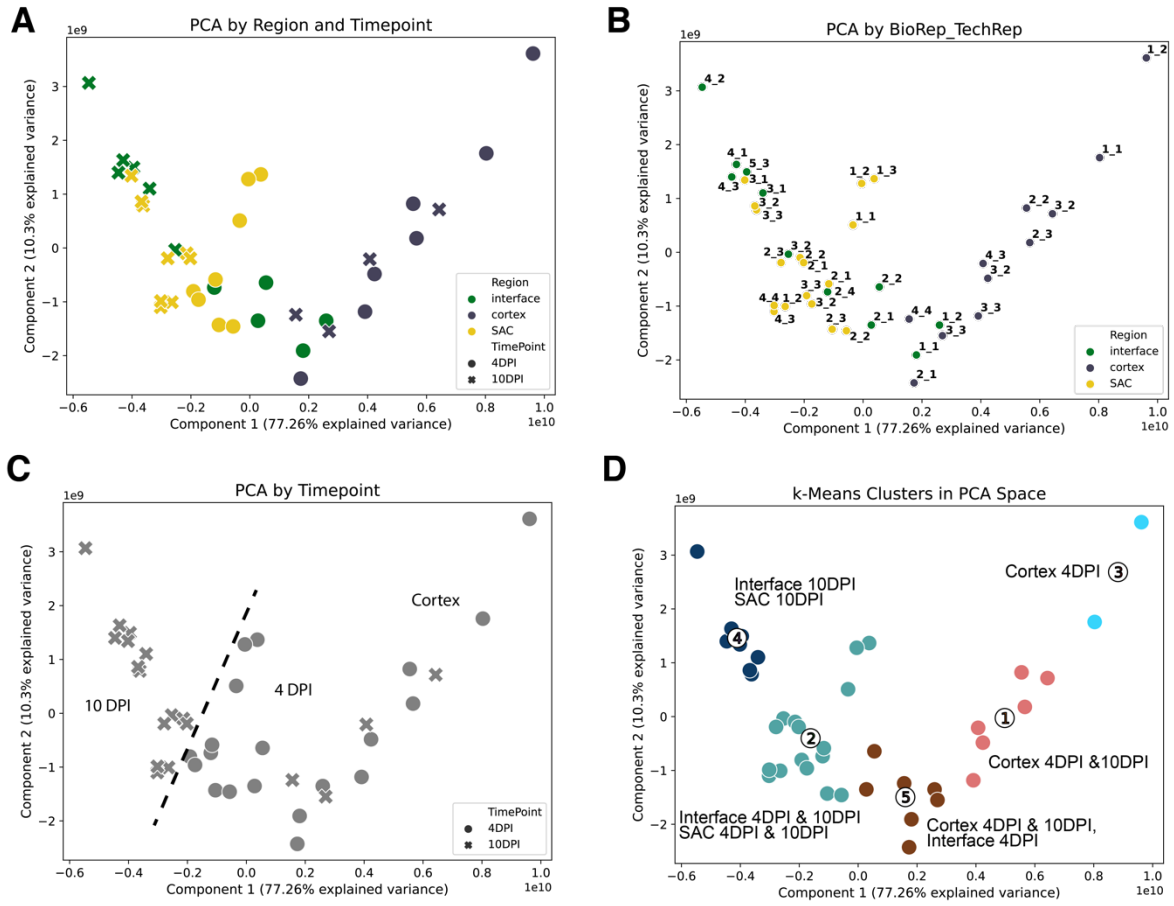

Figure 2-5: Principal Component Analysis (PCA) and k-means clustering results of a proteomic dataset with imputed values of an S. aureus infected murine kidney. A) PCA was performed on protein LFQ intensity values acquired from 3 regions and 2 timepoints. This unsupervised approach separates the SAC and interface (left) from the cortex samples with no visible infection (right). B) Samples also seem to cluster based on biological replicate within the PCA space. C) There is a separation among the samples 4- and 10-days post infection within samples acquired from the region of infection; this separation is not seen from samples acquired from the cortex where there was no visible infection. D) k-means clustering was used to cluster the samples after PCA. k = 5 was determined using silhouette scores as a metric. To aid in interpretation, clusters are labeled by the regions and timepoints from which samples were collected.

The analysis of the cluster centroids revealed that cluster membership is largely driven by the same proteins, such as alpha globin 1, cytoplasmic actin, and beta-globin (Figure 2-6). However, there were some notable new proteins in this list such as several immune-response related factors such as S100-A9 and prothymosin alpha.[98] Although the primary drivers of cluster membership remained the same, the same analytical process applied to the larger imputed dataset provided a broader description of the host-pathogen interface, uncovering additional target proteins that can be further validated.

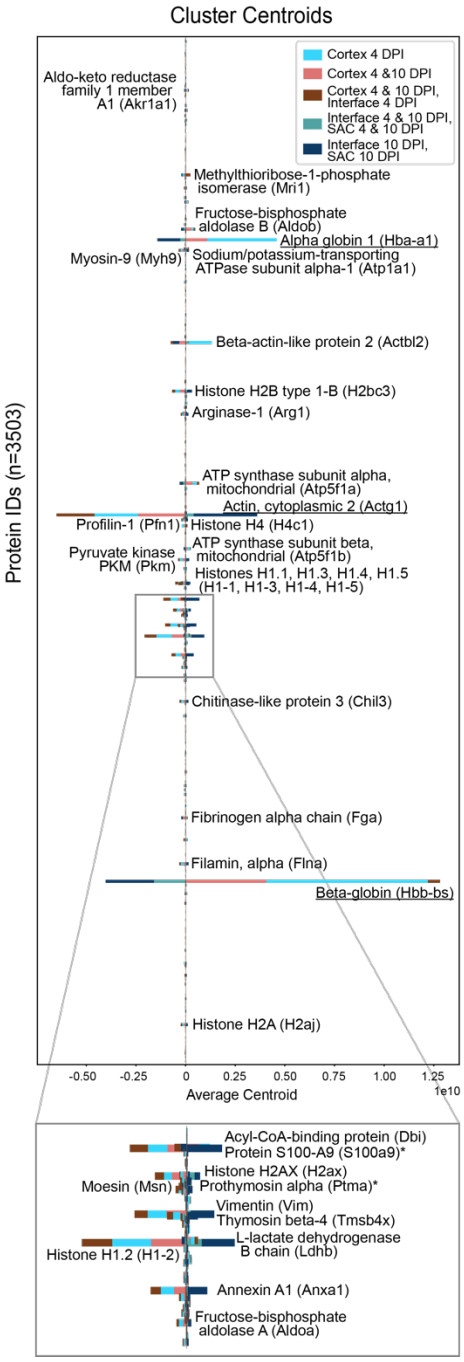

*Figure 2-6: Murine molecular differentiators among regions of S. aureus-infected kidney using the zero-filled dataset. All five cluster centers are overlaid, with 1% of proteins (n=35) with highest absolute values labeled. Underlined = three proteins with highest centroid values; * = proteins involved in the immune response.*

One caveat to this analysis is the interpretation of the PCA pseudo-protein signature, which represents combinations of protein LFQ data that capture as much of the observed variance as possible. However, variation does not always imply biological relevance. For instance, the PCA might be skewed by high-intensity values for proteins that may not be biologically relevant or miss

low-intensity valued proteins that nevertheless may hold biological significance, but whose importance is mathematically hard to discern in the presence of proteins with higher intensity values. Even though normalization across the entire sample set during the LFQ intensity calculation was performed within MaxQuant differences in overall protein intensity values may still affect the final output. This relays a fundamental concern in proteomics, which is the extremes in dynamic range of signal and biological abundance for detected proteins.

Another caveat is that PCA is vulnerable to proteins that may be present in non-Gaussian distributions in the LFQ intensity domain, which goes against a key assumption for PCA.[32] Therefore, it is important to refrain from attempting to over-interpret the pseudo-protein signatures and limit the interpretation to exploring in each cluster only the highest absolute centroid values. In doing so, we only claim to find a focused subset of interesting proteins that merit further investigation, from among the hundreds that were measured over the entire experiment, thereby providing a means of efficiently identifying candidates for future investigation. Another way to mitigate this is to use alternative clustering methods such as hierarchical clustering can also be used to analyze the dimensionality reduced proteomics data. For this analysis, *k*-means was selected due to the ease of interpretation of the cluster centroids, which represent the average protein pattern for each cluster.

In summary, this unsupervised multivariate method provides a way to efficiently analyze highly complex spatially targeted proteomics data and provide an effective way of highlighting a panel of potential drivers of biological differences among regions of interest.

Gene Ontology Analysis

In addition to identifying individual proteins from thousands measured, the functional categories of each protein were assessed to provide additional biological insight. The Protein Analysis Through Evolutionary Relationships (PANTHER) classification system[99] was used for gene ontology analysis of the proteins driving the clustering algorithm. Absolute centroid values were summed across all four clusters, and the top 100 and 175 proteins for the non-imputed and zero-imputed datasets, respectively, with highest accumulated centroid values selected for gene ontology analysis. A broader or more narrow biological interpretation can be performed by selecting more or fewer proteins, respectively. Original LFQ intensity values for the selected proteins were retrieved, and their LFQ intensity was standardized per protein by removing the mean and scaling to unit variance. Standardized protein intensity values were averaged per cluster, proteins with positive values per cluster were extracted, and proteins per cluster were analyzed using PANTHER for a gene ontology analysis, and the resultant protein classes found in each cluster were determined (Figure 2-7).

The gene ontology results indicate that the set of proteins driving the clustering model comprise thirteen protein classes, including cytoskeletal and metabolic processes (Figure 2-7). Panels A - D are sorted from regions distant from the abscess with no visible bacteria present (cortex 4DPI and 10DPI) to those in proximity to abscesses at the later timepoint (interface/SAC 10DPI), and panel E shows three protein classes (cytoskeletal, metabolite interconversion enzyme, and calcium-binding) with distinct changes between regions of infection (interface/SAC) and no infection (cortex/early interface).

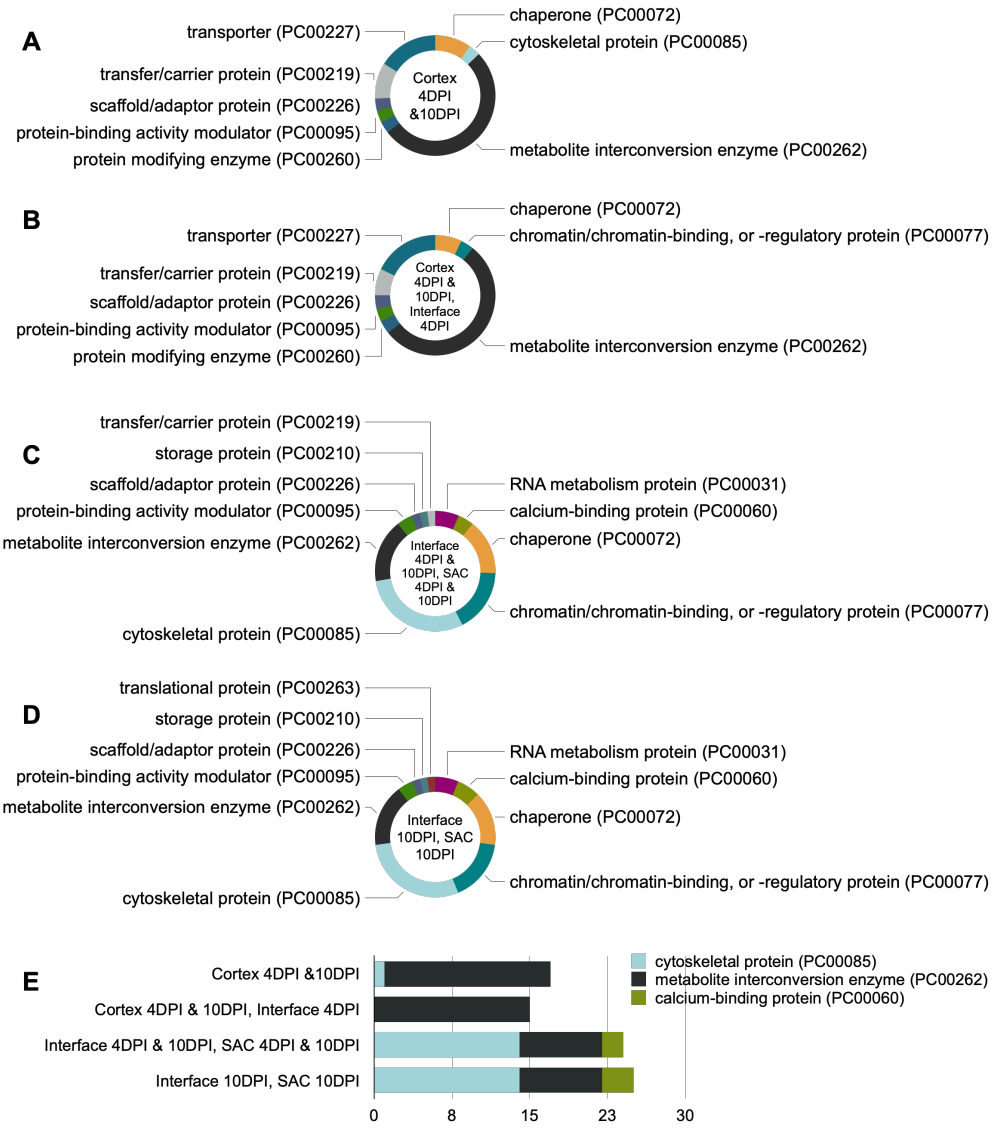

Figure 2-7: Gene ontology analysis. Gene ontology analysis was performed using the 100 proteins with highest accumulated absolute centroid values. The LFQ intensity for these proteins were normalized across all samples and those with values above 0 were analyzed using the PANTHER classification system based on Protein Class. Panels A – D are sorted from no infection (cortex 4DPI & 10DPI) to most infection (interface 10DPI and SAC 10DPI). The total number of proteins in each cluster are as follows: A) 31, B) 28, C) 47, D) 48. Panel E shows three protein classes with differences among regions of infection versus no infection, and the total number of proteins in each class.

A gene ontology analysis was also performed using the larger, zero-filled dataset (Figure 2-8). As with the centroid analysis described previously, protein classes remained largely the same with the addition of defense/immune proteins that were present in clusters with infected samples. There were also additional calcium-binding proteins, such as Calprotectin, a major immune component.

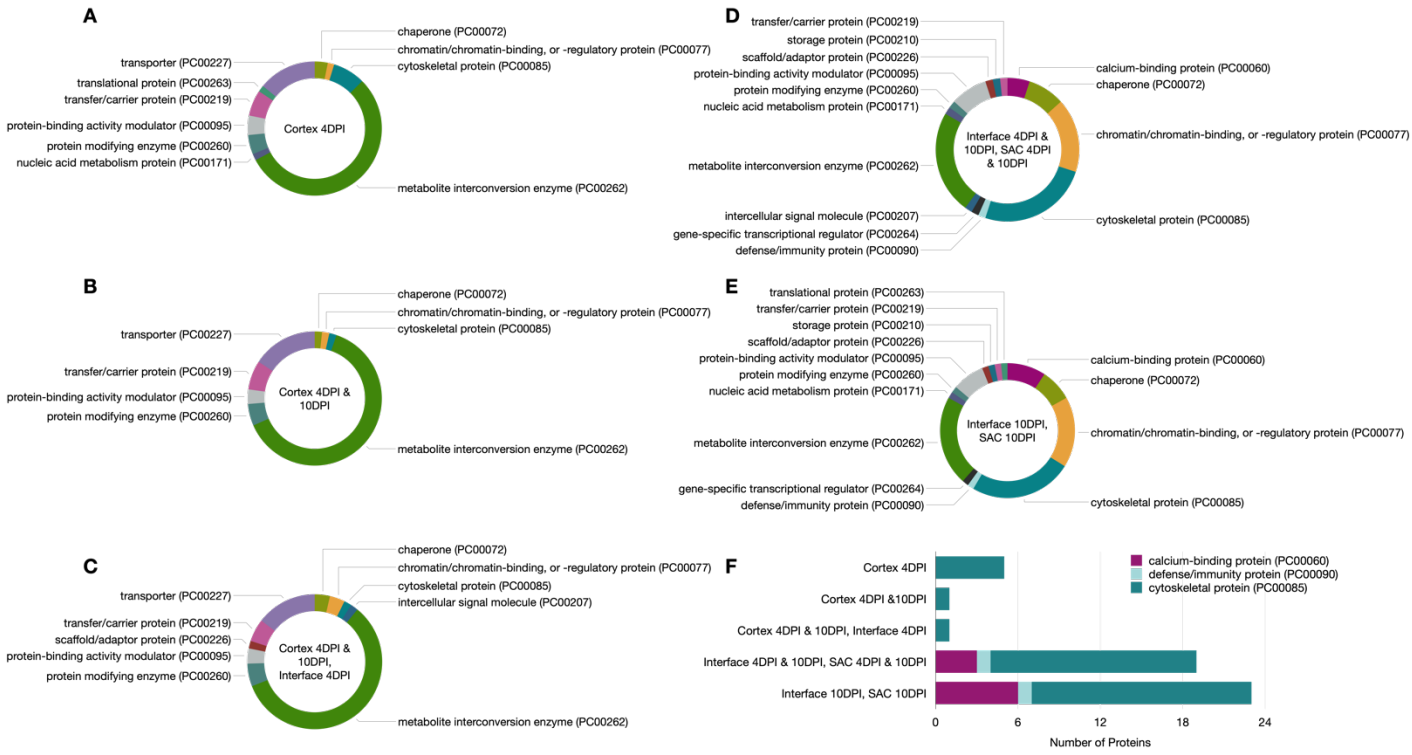

Figure 2-8: Gene ontology analysis using the imputed dataset. The PANTHER classification system was used to perform a gene ontology analysis of the 5% of murine proteins with highest centroid values (n=175). These proteins were classified based on Protein Class. Panels A – E are sorted from no infection (cortex 4DPI) to most infection (interface 10DPI and SAC 10DPI). The total number of proteins in each cluster are as follows: A) 64, B) 57, C) 55, D) 60, E) 65. Panel F shows three protein classes with differences among region of infection and no infection and the total number of proteins in each class.

This analysis identified cytoskeletal proteins that are enriched at the site of infection (abscess and interface), particularly at the later timepoint, indicative of extensive tissue damage resulting from *S. aureus* residing and proliferating within the tissue, as well as subsequent repair and remodeling efforts by the host.[100] In addition, there is an enrichment of established immune factors, such as calcium-binding proteins, comprising different Annexins, which have been recently implicated in the defense against Gram-positive infections.[101–103] Annexins A2 and A3 were increased in two clusters: i) Interface 4DPI & 10DPI, SAC 4DPI & 10DPI and ii) Interface 10DPI, SAC 10DPI. Annexin A5 showed increased abundance only in the Interface 10DPI, SAC 10DPI cluster.

A survey of recent literature suggests that Annexin A2 interacts with staphylococcal clumping factors A and B, facilitating attachment to epithelial cells.[102,103] One study concluded that the binding of Annexin A2 allows *S. aureus* to anchor onto vascular endothelial cells, establishing this host protein as an important factor for initiating staphylococcal interaction with its host.[104] In contrast, little is known about the roles of Annexins A3 and A5 during infection with *S. aureus*. A transcriptomics study revealed that Annexin A3 expression is restricted to neutrophils and is increased in the blood of patients with sepsis.[105] Annexin A5, which was increased in the Interface 10DPI, SAC 10DPI cluster, has been shown to aid survival in a murine sepsis model by inhibiting HMGB1-mediated proinflammation and coagulation.[106] Despite these findings, it is not

clear how Annexins A3 and A5 affect the host-pathogen interplay, particularly in the context of *S. aureus* soft-tissue infections. While our study relies on a relatively small sample size, our data clearly show that Annexins A2, A3, and A5 are highly abundant at the site of infection. The data presented here and previous studies on Annexins allow us to speculate that while A2 may be facilitating staphylococcal anchoring in the tissue, A3 and A5 may confer varying degrees of host protection during staphylococcal infection.

This systems-level analysis of a complex biological model demonstrates the utility of the data generated through this multivariate analysis method and presents the potential for future biology-driven investigation and experimentation. The gene ontology analysis demonstrated here is one of many potential interpretations of the cluster centers. Another method for interpreting the cluster centroids would be to build protein-protein interaction networks using proteins with high accumulated absolute centroid values as seeds. Another caveat within this particular infectious disease model is that the samples chosen in this study belonged to biologically distinct locations with profound protein changes. However, there are multiple opportunities to tune the pipeline to be more robust or sensitive to the protein changes within the study for different data sets with more nuanced protein heterogeneity. There would need to be some prior knowledge about the source of protein variation, but those changes could be used to inform the PCA, the number of k-means clusters, and the approach to cluster interpretation.

Even though this method was applied to spatially targeted proteomics data acquired by microLESA, it can be extended to multi-omics data involving metabolites, lipids, and peptides acquired using other spatially targeted approaches such as liquid extraction surface analysis,[69,71,107] liquid microjunction,[108,109] nanoPOTS,[16–18] tissue punch biopsies,[110,111] laser capture microdissection,[76–79,112] and hydrogel extractions.[113–115]

CONCLUSIONS

A rapid automated unsupervised method for analyzing high-dimensional spatially targeted proteomic data utilizing PCA followed by *k*-means clustering was applied to study soft-tissue *S. aureus* infection in murine kidney. *k*-means clustering results revealed molecular heterogeneity in the abscesses and the interface region between areas of infection and non-infection that goes beyond what can be seen by microscopy alone. Proteins driving the clustering algorithm, and thereby likely to play a role in staphylococcal infection, were extracted from cluster centroids and found to be involved in key metabolic processes and cytoskeletal reorganization. Subsequent gene ontology analysis of proteins with high accumulated absolute centroid values revealed that proteins involved in calcium-dependent, metabolite interconversion, and cytoskeletal processes were enriched in sites of infection, especially at the 10DPI timepoint. These findings collectively demonstrate that this multivariate approach is a powerful method that provides a means of rapidly filtering complex biological data to determine the most relevant species from hundreds to thousands of measured proteins in the form of ranked protein lists and pathway enrichments, thereby providing a systems-level view into complex molecular biological processes.

METHODS

Sampling and Data Acquisition

Data used in the murine case study are stored on the ProteomeXchange Consortium database by the PRIDE118 partner repository with the data set identifier PXD019920.[74] From this original publication, we briefly report the methods used for sample preparation and technical aspects for microLESA and LC-MS/MS (Figure 2-1).[74] Six- to eight-week-old mice were retro-orbitally inoculated with *S. aureus* (strain USA300 LAC) constitutively expressing sfGFP.[74] Infections were allowed to progress until 4 or 10 days post-infection (DPI) before animals were humanely euthanized and kidneys excised for analysis. All animal experiments were approved by the Vanderbilt Medical Center Institutional Animal Care and Use Committee. Kidney sections were cryosectioned into 10 µm thick tissue sections, thaw-mounted onto glass microscope slides, and imaged with autofluorescence microscopy (Carl Zeiss Microscopy, White Plains, NY)) to determine ROIs for microLESA sampling. Trypsin dissolved in ddH2O to a final concentration of 0.048 µg/mL was applied to each ROI using a robotic piezoelectric spotter (sciFLEXARRAYER S3, Princeton, NJ). Slides were then incubated at 37°C for three hours in 300 µL ammonium bicarbonate, and proteolytic peptides were extracted using a TriVersa NanoMate (Advion Inc., Ithaca, NY) with the LESAplusLC modification. To mitigate batch effects, samples were run in a single batch in a randomized order by both region and time point. Samples were stored at -4°C prior to analysis to preserve protein integrity. Once all samples were collected, they were collected and analyzed by liquid chromatography with tandem mass spectrometry (LC-MS/MS) in positive ion mode using an Orbitrap Fusion Tribrid mass spectrometer (Thermo Scientific, San Jose, CA) at 120,000 resolving power at *m/z* 200 with a mass range of *m/z* 400-1600 and an automatic gain control target of $1.0x10^6$.

Data Analysis

Protein identification and quantitation were performed using MaxQuant version 1.6.7[73] as follows (Figure 2-1A). Raw LC-MS/MS files were processed using the label-free quantification method in MaxQuant. Spectra were simultaneously searched against *Mus musculus* and *Staphylococcus aureus* (strain USA300 LAC) reference databases downloaded from UniProt KB[116], and the resultant peptide and subsequent protein identifications include name of species. These labeled identifications can later be used to separate the proteins by species. These were supplemented with the reversed sequences and common contaminants for quality control purposes. Acetyl (protein N-term) and oxidation (M) were set as variable modifications. The option for 'Match between runs' was not used and the LFQ min. ratio count was set to 1. Minimal peptide length was seven amino acids. Peptide and protein false discovery rates (FDRs) were both set at 1%.

The subsequent data analysis was performed on the resultant protein groups file containing label-free quantitation (LFQ) intensity values from MaxQuant, was used for. In this file, each row contains the group of proteins that could be reconstructed from a set of peptides; proteins in each protein group are sorted based on the number of identified peptides in descending order. This protein groups file was analyzed for outliers using a z-score anomaly detection calculation. Briefly, z-scores were calculated based on the number of protein groups identified and samples with z-scores > |2| were excluded. Based on this calculation, 3 samples out of 42 in all were excluded. Proteins identified as "reverse", "only identified by site", or "potential contaminants" were also

removed, as were proteins with fewer than 2 unique peptides identified. As a result of this filtering process and due to the molecular heterogeneity between samples, there are many missing LFQ values in the dataset. For the initial analysis, proteins with missing values in any of the samples were excluded from the subsequent data analysis, resulting in a dataset comprising only 287 proteins (rather than the 3613 protein rows from the start). To also assess broader coverage, a secondary (inclusive) analysis was also conducted, where instead of removal, the missing values were zero-filled and analyzed using the same subsequent data analysis method.

Using Python version 3.7, we applied Scikit-learn's PCA with a randomized solver[90] and generated an array of 39 ranked components (the maximum, given that there are 39 samples). This array of PCA-transformed data (of size 39×39 instead of the original 39×287) was then used for *k*-means clustering using Scikit-learn's KMeans implementation. A range of *k* values from 2 to 15 was tested using silhouette scores[96] as a performance metric[96] to determine the optimal *k* number of clusters. Upon determining the optimal *k* value to be 4, the *k*-means clustering algorithm was deployed to assign cluster membership to each sample; aside from setting the *random_state* parameter to a fixed but randomly selected integer (42) to maintain reproducibility across runs, the default parameters were used. Cluster centroids for each cluster, which represent the average for all points belonging to the cluster, were used for biological interpretation with 10% of proteins (n=29) with highest absolute values labeled. Since the *k*-means clustering was performed on PCA-transformed data, the resultant cluster centroids are in the form of 4 rows (one per cluster) and 39 columns (one for each principal component). To interpret the cluster centroids in terms of the protein groups, we cast the centroids back to the original measurement space by performing matrix multiplication between the centroid table (of size 4 × 39) and the PCA scores table (of size 39 × 287), thereby generating a final matrix of size 4 × 287. For the secondary (inclusive) analysis as a supplement to the original analysis, we also performed the PCA and *k*-means clustering on the full proteomic dataset, zero-filling the missing values, which resulted in a total feature set of 3613 proteins. For this analysis, a *k* of 5 was selected and the resultant cluster centroids were extracted in the same way as described above, with the note that the final centroid matrix was in that case of size 5 × 3613.

The absolute centroid values were summed per cluster and the 100 and 175 proteins for the non-imputed and zero-imputed datasets, respectively, with highest accumulated centroid values were selected for gene ontology analysis. The original LFQ intensity values for those top proteins were extracted for each sample and their intensity was standardized per protein by removing the mean and scaling to unit variance. These standardized protein intensity values were averaged per cluster, and proteins with a standardized intensity greater than zero were selected for gene ontology enrichment analysis, which was performed using the Protein Analysis Through Evolutionary Relationships (PANTHER) classification system (version 16.0) for each set of proteins per cluster.[99] Only murine proteins were used for the gene ontology analysis since PANTHER does not include the *S. aureus* strain USA300 LAC in their databases. The resultant protein classes were summarized. All code for data analysis can be found at https://github.com/kavyasharman/microlesa.

CHAPTER 3 MULTIMODAL MALDI IMS AND CODEX IMMUNOFLUORESCENCE TO ASSESS HOST IMMUNE RESPONSES

OVERVIEW

Spatially targeted mass spectrometry techniques such as matrix-assisted laser desorption/ionization (MALDI) imaging mass spectrometry (IMS) permit label-free analysis of hundreds to thousands of chemical species within a single tissue section and span the fields of proteomics, metabolomics, and lipidomics. However, extracting biologically relevant molecular drivers from such experiments remains a challenge due to the complexity and high dimensionality of the data acquired from each spatial location. Microscopy-based techniques such as co-detection by indexing (CODEX) multiplexed immunofluorescence (MxIF) and histological staining make it possible to demarcate a number of cell types and functional tissue units (FTU's), providing a means of contextualizing spatially targeted mass spectrometry data. In this work, a series of segmentation workflows were assessed, leading to the development of a customized multivariate *k*-means clustering approach was developed to generate segmentation masks, which were used to probe MALDI IMS data to explore the cellular and lipidomic composition of an *S. aureus*-infected murine kidney. Results uncovered lipidomic heterogeneity among abscessed regions and non-abscessed regions, with specific lipids localizing to each region, demonstrating the utility of this integrated workflow for understanding the host immune response.

INTRODUCTION

Imaging mass spectrometry (IMS) allows for molecular interrogation of tissue while preserving spatial integrity.[21,22] IMS studies span the fields of proteomics, metabolomics, and lipidomics, permitting label-free characterization of tens to thousands of chemical species within a single experiment.[23–27] However, extracting biologically relevant molecular drivers from spatially targeted mass spectrometry experiments remains a challenge due to the complexity of the dataset, with a full mass spectrum comprising tens to thousands of molecular measurements gathered from each spatial location. One way to contextualize IMS data is to supplement it with registered autofluorescence (AF),[7,117] which provides gross anatomical information for automated image registration, or co-detection by indexing (CODEX) multiplexed immunofluorescence (MxIF) microscopy, [10–12] which labels cell-specific antigens with antibody-bound fluorescent markers.

Traditionally, this gross anatomical information can be provided by an expert pathologist or domain expert trained to identify cell types and functional tissue units. However, as this is a manual process, it is typically performed on small portions of a tissue section. With large whole-slide images (WSIs), manual annotation becomes cost and time-prohibitive due to the high number of tissue substructures, which brings with it the challenge of avoiding human bias and human drift in accuracy as more and more structures are annotated. Through the development of computational approaches to whole slide images, segmentation can be applied across the entire tissue section. There have been myriad approaches in the microscopy and imaging fields to build automated segmentation techniques and software. Many of these segmentation techniques have been developed for the single cell level, relying on first identifying cell centers and then identifying cell borders.[118–120] Others rely on applying deep learning approaches[121] sometimes requiring interactive training[122,123] to build a model that can then be deployed to generate segmentation

masks. Still others rely on integrating existing methods to generate an analytical toolbox to enable exploration of cell phenotypes.[124] However, most, if not all, of these approaches are not as well suited for exploring larger cellular substructures and functional tissue units (FTUs). For instance, within the kidney, FTUs such as glomeruli or tubules are comprised of groups of cells that are not only much larger in size but can be appear to be disconnected since they tend to be larger than the thickness of a standard serial section and are therefore only seen in fragments. Although the aforementioned methods perform well for nuclear and single cell segmentation, they do not perform well for these larger FTUs. This challenge necessitates the development of segmentation pipelines that are better suited for multi-cellular units.

To address these challenges, a series of segmentation techniques were tested, leading to the development of a customized unsupervised multivariate segmentation workflow. This was applied to study the host immune response in a *Staphylococcus aureus*-infected murine kidney. Abscess formation is a hallmark of *S. aureus* infection and disease progression. Once thought to be a static lesion, the abscess is now understood to be a dynamic microenvironment of staphylococcal microcolony surrounded by layers of staphylococcal cells releasing microbial factors, host immune cells releasing antimicrobial factors, and living and dead host tissue cells.[39,40,125] Within this molecularly heterogeneous interface, lipids have been specifically implicated in a number of host immune responses[126–129] as well as key biological functions.[130,131] Furthermore, lipids may also be implicated in antibiotic resistance of *S. aureus*.[132] In sum, there is much that remains to be resolved in terms of specific molecular factors that are involved in the development and progression of these abscesses. Here, we leverage multimodal MALDI IMS and CODEX MxIF to help elucidate the lipidomic landscape of staphylococcal abscesses and the host immune response.

RESULTS & DISCUSSION

The staphylococcal host-pathogen interface within a soft-tissue infection is a complex molecularly heterogeneous environment. Before developing a segmentation technique to generate masks of cell types and functional tissue units on the basis of CODEX MxIF data, we first obtained a normal murine kidney with no infection present to evaluate a series of segmentation techniques.

Image segmentation methods were evaluated for their potential in segmenting substructures of a bacterial infection. Watershed segmentation and intensity thresholding was used to identify and segment regions of the bacteria. The automated segmentation algorithm was built using QuPath,[34] an open-source software for digital pathology and whole slide image analysis, to automatically segment. This process was scripted in Java as a QuPath macro to enable automated segmentation across the WSI. The first part of the script involved tiling the WSI. Overlapping tiles are generated to ensure annotations that may fall on the border between two tiles are captured. This tiling process was found to be most computationally efficient when coupled with the segmentation. As a result, the method entails generating a tile, exporting the tile into ImageJ, segmenting, and exporting the resultant mask back to QuPath. A detailed workflow for segmentation for each tile was built (Figure 3-1) that involved an initial preprocessing step converting the image to RGB and applying a Gaussian blur for denoising the image. A watershed algorithm was applied to the values in the image to determine the local maxima. Concurrently, the same filtered tile is thresholded using an intensity threshold and an intersection of the maxima image and thresholded image is calculated. Holes are filled to ensure entire FTUs are being captured before creating a selection and exporting the resultant annotations back to QuPath. Once the annotations are created, they can

be exported as vector-based segmentation masks as .txt files with x and y coordinates that can be applied to the registered IMS data.

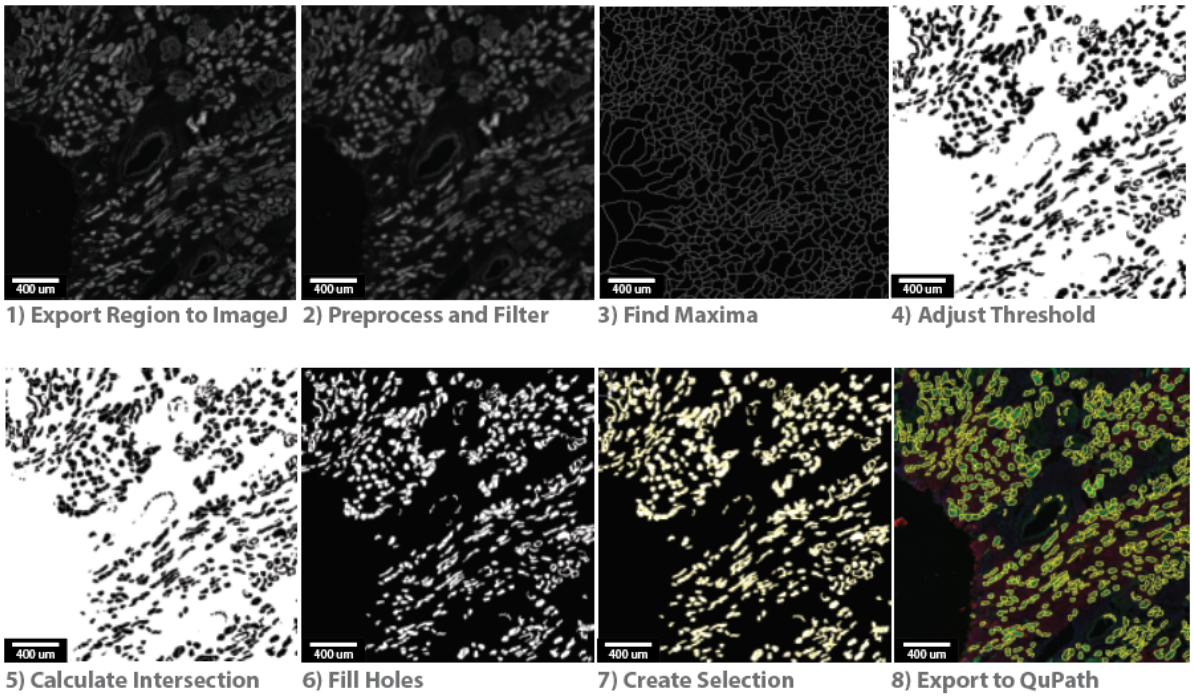

Figure 3-1: Watershed and intensity-based segmentation approach on a single tile. Individual tiles are exported from QuPath, an open source digital pathology image analysis software, to ImageJ and processed for single channel intensity thresholds and watershed maxima, resulting in a binary mask output highlighting unique macro structures in the tissue. This binary mask is split into individual segments as QuPath annotations and exported for registration with pre-processed IMS data.

This tiling approach was applied to a single channel of a WSI acquired on a normal murine kidney (Figure 3-2). The overlapping tiling approach ensured that any FTUs that were on the border of a tile were captured fully. Although the results show successful segmentation of the renal tubules, closer analysis reveals that some of the segmentations split tubules into multiple ROIs.

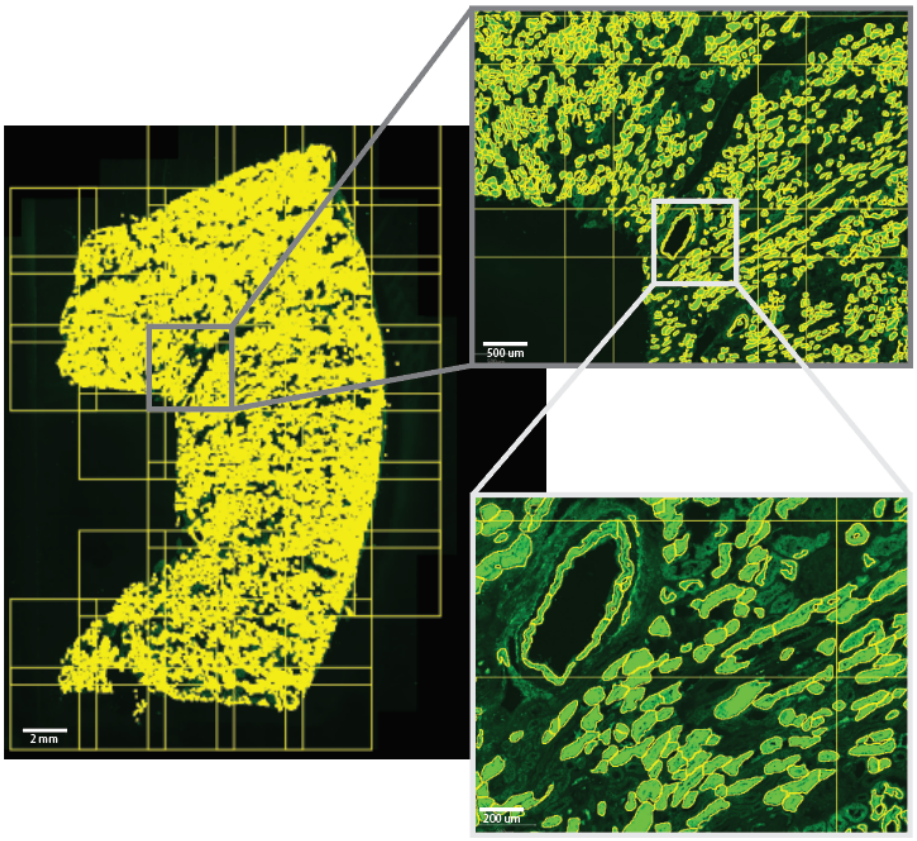

Figure 3-2: Whole-slide segmentation of murine kidney on the AQP1/AF 488 (FITC) channel, which marks proximal tubules. Images are tiled and segmented sequentially to improve processing speed.

This challenge of split segmentations is consistent across multiple FTUs in the murine kidney as well (Figure 3-3). For instance, upon looking at the original unsegmented four-channel MxIF image (Figure 3-3a), it is evident that there are varying sizes of tubules. However, upon performing segmentation on the proximal tubules (Figure 3-3b), thick limb (Figure 3-3c), and collecting ducts and proximal tubules (Figure 3-3d), it is evident that many of the FTUs have been split across many segmentations. Nuclei segmentation (Figure 3-3e,f) is the only FTU that performs well with this type of watershed and intensity-based segmentation, largely due to the uniformity of the IF signal as well as the relatively similar size of each nuclei. Additionally, due to non-specific markers such as biotinylated DBA/NeutrAvidin 650, which marks both collecting ducts and proximal tubules, identifying specific FTUs remained a challenge. A final challenge with this type of segmentation approach was the manual determination of optimal hyperparameters for color thresholding levels and noise tolerance. Although selecting these for a single WSI is feasible, applying this segmentation as a standardized workflow across multiple WSI microscopy images with different relative intensity levels is not possible without optimizing hyperparameters for each image, thereby hindering reproducibility. It was therefore concluded that although the tiling approach that was developed worked well to uniformly segment WSIs using a watershed and intensity-based approach, it was not well-suited for our goal of segmenting distinct FTUs for the purpose of integrating them with IMS and generating FTU-specific molecular signatures.

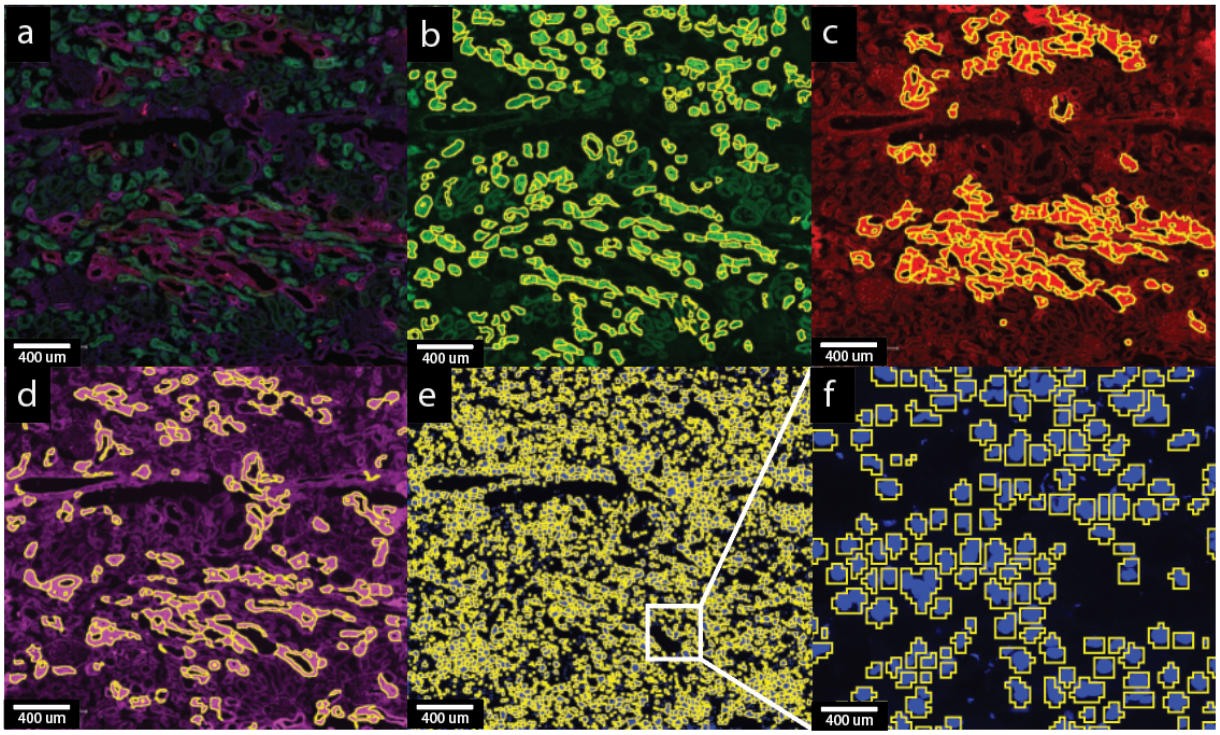

Figure 3-3: Multi-channel segmentation on an IF image of normal murine kidney. A) original unsegmented four-channel MxIF image. B) AQP1/AF 488 (FITC), which marks proximal tubules. C) THP/Cy3 (Texas Red) stain, which marks thick limb. D) Biotinylated DBA/NeutrAvidin 650 (Cy5), which marks collecting ducts and proximal tubules. E) Hoechst (DAPI), which marks nuclei. F) a higher resolution of Hoechst (DAPI) stain showing segmentation of individual nuclei.

To address the challenges of reproducibility, scaling, and disparate segmentation masks of single FTUs, a multichannel segmentation workflow comprising singular value decomposition (SVD) followed by *k*-means clustering was developed. SVD is similar to PCA in that it addresses the "curse of dimensionality," or the challenges in analyzing high dimensional data, and groups correlated and anticorrelated features into a set of orthogonal components which can then be clustered using a method such as *k*-means. By grouping features using SVD before *k*-means, lower-amplitude differences among features can be captured, thereby improving *k*-means clustering capabilities.

Another challenge with processing large WSI images is that it is difficult to perform analyses on the entire image at once due to limited computational power. One solution is to apply parallel processing, which allows the user to work with datasets that are larger than the current working memory. This was accomplished by first saving the .ome.tiff microscopy file as a Zarr file and then loading the image using Dask[133] arrays. Dask arrays are built by storing the full Zarr file into a series of smaller "chunks" that can be loaded separately and processed in parallel, allowing for more efficient processing of the WSI. This is especially useful for large datasets without losing quality or having to use a tiling approach, which can be sensitive to large-scale artifacts such as gradients across the WSI.

An additional customization we developed in this workflow was the implementation of mini-batch *k*-means clustering[134] from the Sci-kit learn Python programming suite rather than the standard *k*-means clustering implementation. This approach uses mini-batches to reduce the overall computation time, with the size of the mini-batch being a user-specified hyper-parameter. We found that 10% of the full data size provided results similar to standard *k*-means clustering with a processing time that was orders of magnitude faster.

The final customization was the inclusion of a background subtraction step to eliminate intensity signals from off-tissue pixels which introduced noise to the overall clustering results. To do so, the DAPI channel staining nuclei was selected and a series of thresholding techniques from the OpenCV image processing library was tested. It was determined that a global thresholding technique provided the most comprehensive background subtraction; a morphological transformation to smoothen the image and fill holes was applied to ensure all pixels that were on-tissue would be included, and the resultant mask was used to select only on-tissue pixels from the full multi-channel dataset.

With all these optimizations in mind, the unsupervised multivariate segmentation method was deployed onto an MxIF multi-channel image of normal human kidney (Figure 3-4). Mini-batch *k*-means clustering was performed on a random 10% subset of on-tissue pixels with *k* values ranging from 2 to 70. A within-cluster sum of squares (WCSS) score was calculated for each clustering result, and it was determined that a *k* value of 20 was optimal. The results of the clustering with *k=20* was displayed as a multichannel image (Figure 3-4A). We found that clusters 1 and 2 correlated well with glomeruli (red) and tubules (green), respectively (Figure 3-4B). This was especially interesting because the original MxIF marker panel did not contain any glomeruli-specific antibodies; just epithelial markers that stained both glomeruli and tubules. However, in using a multivariate *k*-means clustering approach, we were able to discern glomeruli specifically and create a segmentation mask for them. We also observed clusters correlated with medullary arrays (cluster 5 and 6) and inner and outer regions of varying tubules (clusters 2, 9, and 10).

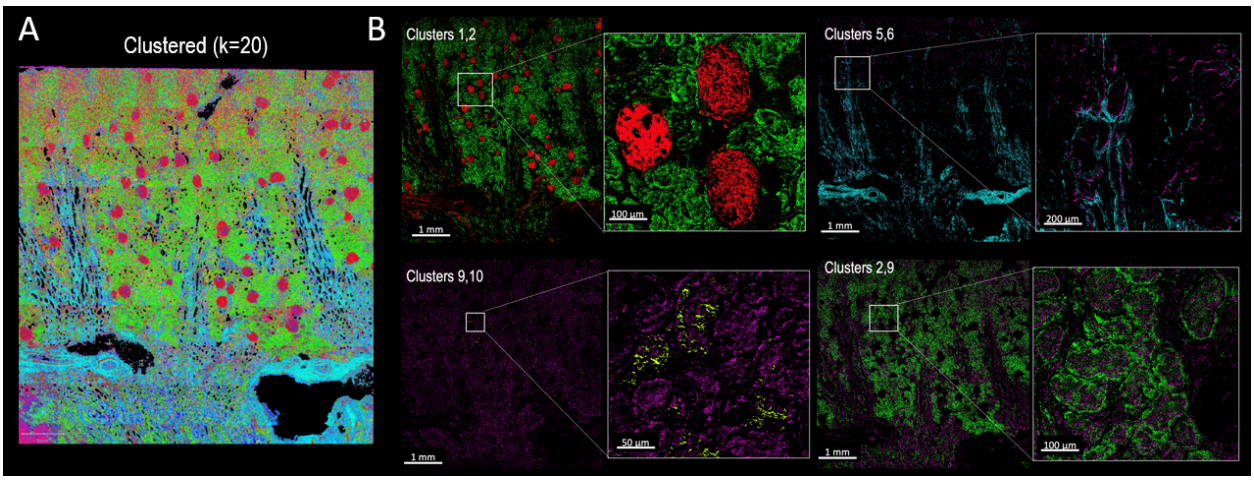

Figure 3-4: Clustered MxIF image of normal human kidney. Clusters 1,2 correlate with glomeruli (red) and tubules (green). Clusters 5,6 correlated with medullary arrays (blue) and epithelial cells (pink). Clusters 9,10 are both correlated with inner (yellow) and outer (pink) regions of tubules. Clusters 2,9 are also correlated with inner (pink) and outer (green) regions of tubules. However, the differences among clusters 2, 9, and 10 seem to correlate with different regions of tubules, potentially highlighting sub-tubular differences.

Having tested this customized segmentation pipeline for highly multiplexed IF data, we applied the same workflow to an *S. aureus*-infected murine kidney. Briefly, mice were inoculated with *S. aureus* (strain Newman) and sacrificed seven days post-infection. Murine kidneys were harvested, cryosectioned at 10µm thickness, and thaw mounted onto coverslips. AF microscopy was acquired on all sections, serving as the basis for multimodal image registration. Sections were analyzed with MALDI IMS to generate pixel-wise molecular data and serial sections were stained against a 17-marker panel comprising immune and renal antigens for CODEX MxIF.

Following image acquisition and registration, the CODEX MxIF imaging data was analyzed using the customized segmentation pipeline. A *k* of 17 was determined optimal, once again using WCSS scores as a metric. In doing so, we observed clusters that matched known cell types and FTUs. For instance, there were four clusters that highlighted areas of abscess and necrosis (Figure 1-5A). Additional clusters revealed different immune and renal regions (Figure 3-5B).

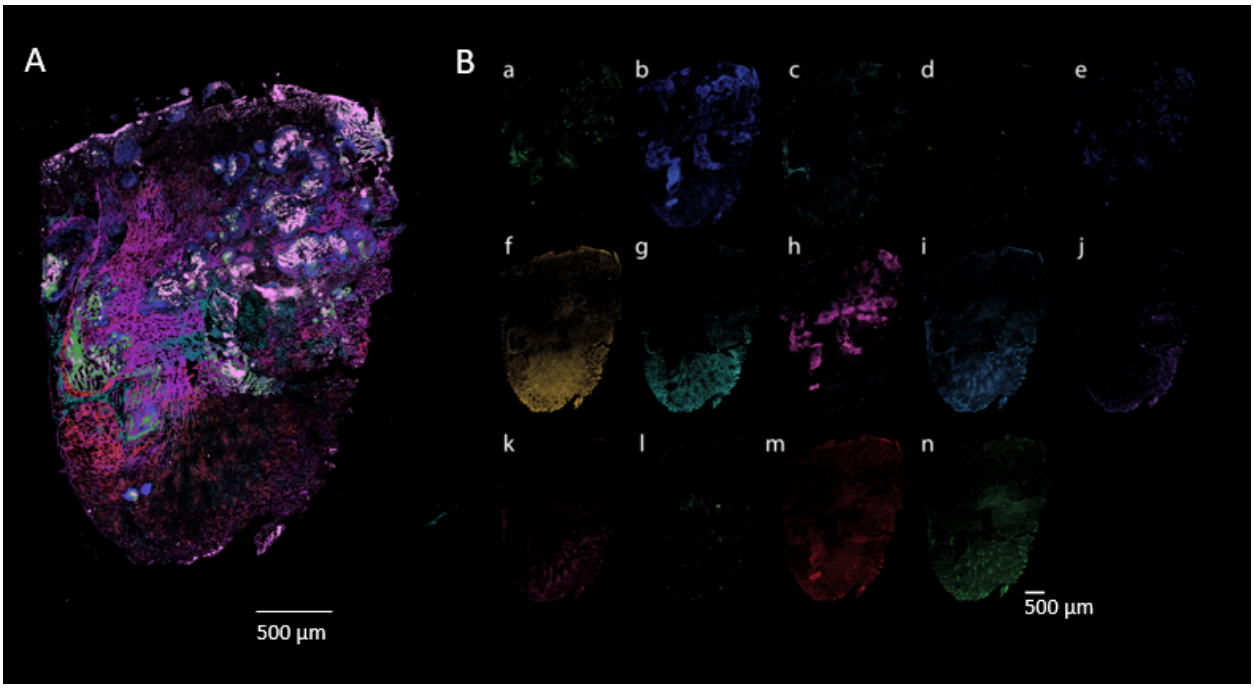

Figure 3-5: k-*means clustering results on* an S. aureus-*infected murine kidney. A) Composite* k-*means clustering image highlights areas of abscess using clusters a, b, h, and j. B) Individual clusters which can be correlated to different immune and renal regions.*

Average mass spectra for the major regions within the kidney and abscessed regions were extracted from the IMS datasets using the pixel coordinates from the *k*-means clustering results (Figure 1-6). In doing so, we detected lipidomic heterogeneity between the abscessed regions. For instance, abscess rich regions displayed $[SM(d34:1)+H]^+$ and $[PC(P\text{-}34:0)+H]^+$, while healthy, non-abscessed regions contained $[PC(36:1)+H]^+$ and $[PC(34:1)+H]^+$ (Figure 3-6A-E).

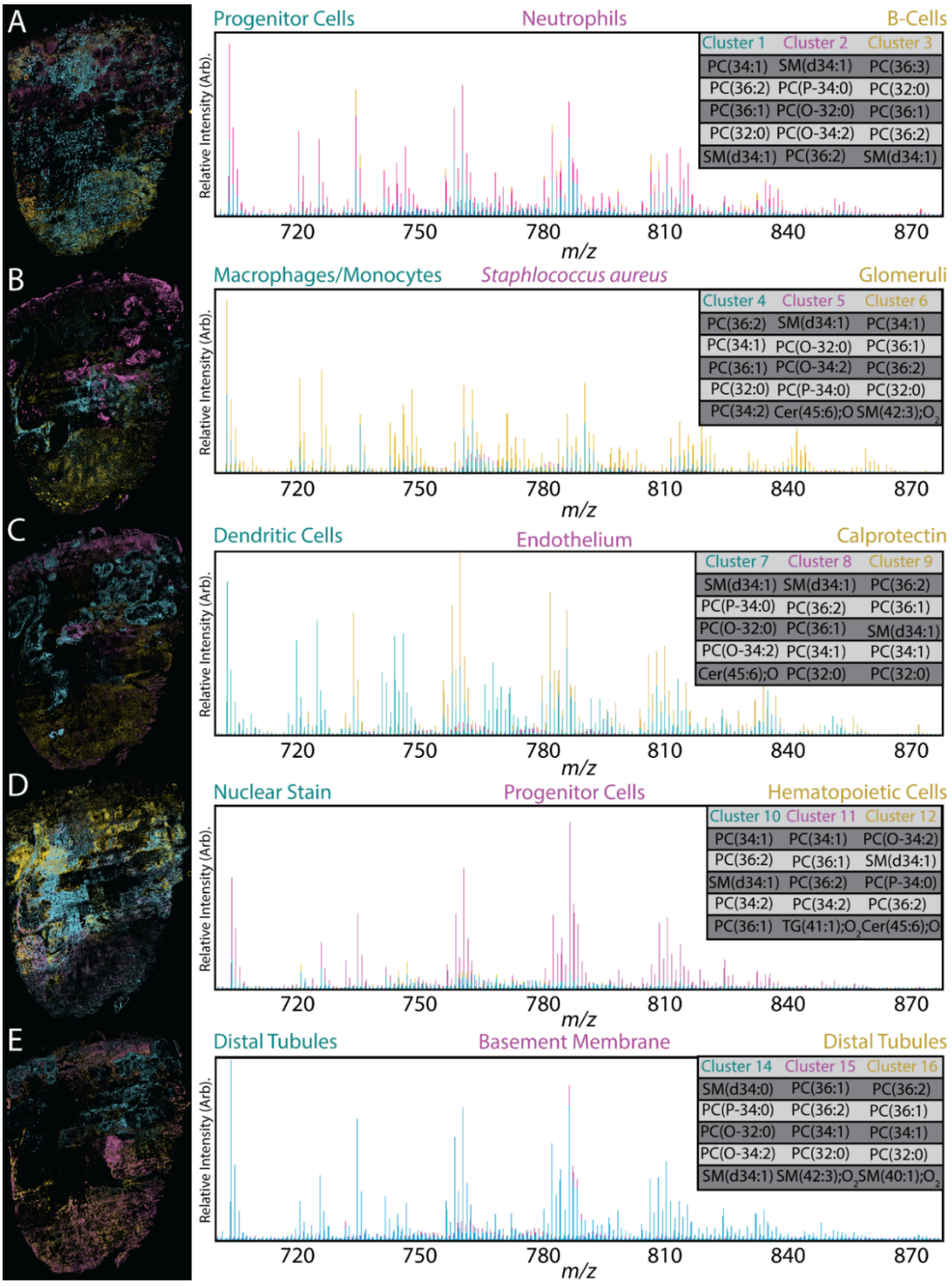

Figure 3-6: Summary of k-means clustering data and extracted mass spectra. A) Overlay of clusters with a high abundance of progenitor cells, neutrophils, B-cells with the top spectral markers for each cluster. B-E) Similar analyses can be seen for each stained cell class. Progenitor cells and distal tubules were represented by two clusters potentially indicating differential response to infection.

CONCLUSIONS

The host immune response within an *S. aureus*-infected murine kidney is complex and remains relatively uncharacterized. Approaches such as MALDI IMS can provide tens to thousands of molecular measurements across the tissue but require integration with orthogonal approaches, such as CODEX MxIF, to correlate the spatially targeted molecular measurements with regions of healthy and infected tissue. In this work, a series of segmentation workflows were assessed to address the challenges of high dimensionality of MxIF imaging and large dataset sizes, leading to the development of a customized unsupervised multivariate *k*-means approach to segment CODEX MxIF data. The segmented imaging data was then used to extract region-specific MALDI IMS signatures, uncovering rich molecular profiles of specific cellular regions within the infected kidney. Biocomputational methods such as those developed here are critical for exploring molecular and cellular diversity and will be required as we move towards more multimodal approaches to study health and disease.

METHODS

Materials:

1,5-Diaminonaphthalene (DAN) was purchased from Sigma-Aldrich Chemical Co. (St. Louis, MO, USA). HPLC-grade tetrahydrofuran (THF), acetonitrile, and methanol were purchased from Fisher Scientific (Pittsburgh, PA, USA).

Murine Infection:

In preparation for murine infections, bacteria were streaked from freezer culture on trypticase soy agar with antibiotics, as required. Isolated colonies were used to prepare overnight cultures in 5 mL of trypticase soy broth (TSB). After overnight growth, bacteria were sub-cultured 1:100 in fresh TSB and grown until mid-to-late log phase (2-3 h; OD 600 nm ~2-2.5). Cells were pelleted by centrifugation at 7,000 rpm for 6 min and washed with phosphate buffered saline (PBS) three times. After the final wash, cells were resuspended to an OD 600 nm of approximately 0.4 in PBS (~1-2 x 108 CFU/mL). The inoculum was determined by serial dilution in PBS and plating to TSA.

Prior to infection, female 6-8-week-old BALB/c mice, purchased from The Jackson Laboratory, were anesthetized by intraperitoneal injection with 125–250 mg/kg of 2,2,2-tribromoethanol. Mice were infected retro-orbitally with 100 µL of the prepared staphylococcal cells (~1-2 x 107 CFU). At 4, 7, or 10 days post infection mice were humanely euthanized by CO_2 inhalation. The kidneys were harvested and immediately frozen on dry ice. Animals were used and handled in accordance with protocols approved by the Vanderbilt University Institutional Animal Care and Use Committee (IACUC) and in compliance with NIH guidelines, the Animal Welfare Act, and US Federal law.

Sample Preparation:

Infected kidney tissue was cryosectioned to a 10 µm thickness, thaw mounted onto indium tin-oxide (ITO) coated glass slides (Delta Technologies, Loveland, CO, USA) for IMS analysis or poly-L-lysine coated glass cover slides for CODEX IF analysis and returned to ~20 °C within a vacuum desiccator. Autofluorescence microscopy images for the IMS was acquired using EGFP, DAPI, and DsRed filters on a Zeiss AxioScan Z1 slide scanner (Carl Zeiss Microscopy GmbH,

Oberkochen, Germany) prior to matrix application.[7,117] IMS samples were coated with a 20 mg/mL solution of DAN dissolved in THF using an HTX TM Sprayer (HTX Technologies, LLC, Chapel Hill, NC, USA) yielding a 1.67 mg/cm^2 coating (0.05 mL/hr, 4 passes, 40 °C spray nozzle). Tissue samples were imaged immediately after matrix deposition before undergoing hematoxylin and eosin histological staining.[135]

MALDI timsTOF IMS:

MALDI IMS was performed on a prototype Bruker timsTOF pro MS system (Bruker Daltonics, Bremen, Germany) in quadrupole-time of flight (qTOF) only analysis mode. The qTOF ion images were collected in positive ion mode at 10 µm pixel size with the beam scan set to 8 µm^2 using 200 laser shots per pixel and 18.6% laser power (30% global attenuator and 62% local laser power) at 10 kHz. Data were collected from *m/z* 50 – 2000 for lipid analysis. All qTOF mode imaging data were visualized using SCiLS Lab Version 2019 (Bruker Daltonics, Bremen, Germany). Lipids were identified using a combination of mass accuracy (≤ 3 ppm) and LIPIDMAPS[136–138] database searching.

CODEX Multiplexed Immunofluorescence:

Samples for CODEX IF were prepared according to the manufacturer's protocols (Akoya Biosciences, Marlborough, MA) and as previously described.[139] Antibodies were conjugated as previously described[139] and diluted to 1:200. CODEX multiplexed immunofluorescence images were acquired on a Zeiss Axio Observer (White Plains, NY) using a 20x objective (324 nm/pixel), LED stack (specifically, 385 nm, 469 nm, 555 nm, and 631 nm), z stack (11 slices at 1.5 µm spacing), and tiling functions. Instrument autofocus was used to focus the imaging area. CODEX IF images were processed using MAUVE software (Akoya Biosciences) which performed the neighborhood analysis automatically with 10 to 30 µm spatial distances.

Data processing:

Data was processed using a combination of commercial and in-house software. Microscopy imaging data was processed in Python. First, images stored as directory-store Zarr[140] arrays were computationally mapped using the Dask[133] for distributed computing. After initial import, the data were pre-processed by removing the mean of each channel and scaling to unit variance using the following calculation: $z=(x-u)s$ (x, sample value, u=mean of all samples with that particular channel, s=associated standard deviation). This pre-processing method was chosen to mitigate any channel intensity artifacts due to differences in microscopy acquisition. We then performed dimensionality reduction by applying a compressed singular value decomposition (SVD) using the SciPy implementation in Python with the 'gesvd' lapack driver.

The *k*-means clustering model was built using a 10% randomly selected subset of the reconstructed matrix following SVD (48,438,400 pixels). A grid search for the ideal number of *k* clusters ranging from 2 to 100 was performed using Scikit-learn's implementation of the *k*-means clustering algorithm. The grid search results were evaluated using a within-cluster-sum-of-squares (WCSS) score, which calculates the averaged squared Euclidean distance of all points within a cluster to the cluster centroid. Using the WCSS score as a metric, a *k* of 17 was selected for the final clustering. The *k*-means clustering model was built on a 10% randomly selected subset of the data and used to predict on the full dataset.

MALDI IMS data preprocessing

MALDI IMS data was processed using in-house Python scripts. The processing creates a mean mass spectrum from the raw IMS data and selects peaks from spectrum with a signal-to-noise ratio of 3 with an intensity threshold of 10 a.u. to remove noise spikes in the data. Peak selection is estimated with full width at half-mass and this *m/z* window is extracted from every mass spectrum to create a dense matrix of IMS coordinates x peak intensities.

MALDI IMS and microscopy image registration

After the MALDI IMS experiment, the MALDI IMS slide was scanned with an Zeiss Axio.Scan.Z1 using an eGFP filter with the matrix layer still on the top of the sample surface. This captures the post-IMS AF image that indicates the position of each laser ablation mark across the tissue surface. Following previously described approaches, the theoretical coordinate of each IMS pixel was extracted from the IMS metadata into an IMS pixel map and the post-AF image was registered to the IMS pixel map by selecting 5 fiducials. The corresponding fiducials here are IMS pixel and its laser ablation mark as imaged by microscopy. This creates an exact registration of the IMS pixel its origin in microscopy coordinates.

After alignment of the postAF image to IMS, the pre-acquisition AF image is automatically registered to the registered postAF image using an in-house registration library that wraps the *elastix* registration tools. Then the CODEX image is also automated registered with the in-house tool to the pre-acquisition AF image previously aligned to post AF image. After alignment of CODEX to pre-acquisition AF, CODEX is registered to the IMS data.

MALDI IMS data analysis

With the CODEX data previously registered to the IMS data, CODEX *k*-means clusters were also aligned with IMS data. As IMS data was sampled at 10 μm/px and the CODEX at 0.5 μm/px, a frequency table of each CODEX cluster per each IMS pixel was computed (*i.e.*, position x101,y101 has n pixels per k cluster). These frequencies were normalized by the IMS pixel area in microns (400 CODEX pixels per IMS pixel) and these values were used to weight the mean spectrum calculation for each *k* cluster. After computing the cluster mean spectra, the overall mean spectrum was subtracted from each to find the most intense signals per cluster. The top markers were taken to be those with the highest signal in the respective cluster's mean spectrum after removing the overall mean.

CHAPTER 4 AUGMENTING DIGITAL PATHOLOGY WHOLE-SLIDE IMAGES WITH MALDI IMS-DERIVED MOLECULAR CONTOUR MAPS

OVERVIEW

Imaging mass spectrometry (IMS) provides spatially informed molecular profiles from tissue samples. Routine visualization of these data are in the form of heat maps; however, interpreting these data and integrating them with known histopathological stained microscopy can pose a challenge due to the complexity of the data from each whole slide image and the lack of methods to integrate the two into a single two-dimensional representation. Here, we develop a contour mapping approach to visualize the ion intensity data from IMS and project the results onto stained microscopy images to study the host-pathogen interface of a *Staphylococcus aureus*-infected murine kidney. Univariate analysis of the IMS data revealed lipids colocalizing with staphylococcal abscesses and contour maps were generated, revealing the two-dimensional lipid distribution within and around the abscess, as well as a quantitative indication of the spatial rate of change of the ion intensity. A multivariate non-negative matrix factorization approach was applied to reduce the dimensionality of the full IMS dataset, generating a subset of thirteen representative images from a full dataset of 440. Visualizing these results as contour maps overlaid onto stained microscopy revealed distinct molecular profiles of the major abscess and surrounding immune response. This workflow also allowed for a molecular visualization of the transition zone at the host-pathogen interface, providing more information about the spatial molecular dynamics than histopathological staining alone. In summary, we developed an innovative visualization strategy using contour maps to project ion intensity data onto high-resolution stained microscopy, thereby providing augmented visualization of the molecular composition of an *S. aureus*-infected kidney.

INTRODUCTION

Matrix Assisted Laser Desorption/Ionization (MALDI) imaging mass spectrometry (IMS) enables concurrent label-free analysis of hundreds to thousands of chemical species within a single tissue section in a single experiment, reporting the spatial distributions of proteins, metabolites, or lipids in the form of ion images.[141–145] However, interpretation of the multitude of molecular images from IMS experiments can be challenging.[22] Often, to aid human interpretation, the molecular images of IMS must be contextualized by other well-characterized and established microscopy modalities, such as histological color stains or immunohistochemistry.[146]

Experiments that combine MALDI IMS and microscopy have become routine and most IMS software, open-source or commercial, has some support for simultaneous visualization of microscopy and IMS images. For multimodal molecular imaging studies where not only spatial co-registration, but also computational integration of the content of the source modalities into a single image type or result is the goal, advanced machine learning (ML) methods have been developed to mathematically integrate the observations in IMS measurements with the observations reported by microscopy into a combined form (e.g. data-driven image fusion,[147]

interpretable ML-based marker discovery[33]). However, for multimodal molecular imaging studies where human interpretation by domain experts is the objective, computational integration of multimodal content is not necessarily the end goal. In those scenarios, it is often important to provide a human-digestible representation that on the one hand takes spatial mapping of the different source modalities into the same coordinate system as far as possible, to reduce the cognitive bandwidth involved in cross-modal spatial mapping of observations, and that on the other hand still allows the content of the source modalities to be viewed and considered separately. Maintaining the ability to view the original microscopy content, textures, and coloring can be important for domain experts to be able to recognize the structures and cues they have been trained on and are familiar with. The contour approach described here fits in the latter category.

Most current software allow the user to visualize ion images side-by-side or to overlay them with microscopy images by registering the images and changing each's opacity. These views offer qualitative insight for cross-modality interpretation, but there remains opportunity for novel visualizations to bring out other different aspects more clearly or for these depictions to be tailored towards particular applications. In this work, we are particularly interested in making high-dimensional data in combination with microscopy easier to interpret for humans and in more clearly delineating important spatial areas of molecular change and making their correspondence to specific microscopic areas more accessible for domain experts.

Here, we develop the use of contour maps for combining the untargeted molecular information from IMS with biologically informative microscopy. Contour maps have historically been used in geography to depict land structures and elevations. In these maps, the proximity of the contour lines represents the change in altitude, allowing the viewer to visualize in a two-dimensional space the localized height of geographical objects such as mountains, as well as the rate of altitude change across space. Within the biomedical community, contour maps have been used to depict risk of disease recurrence,[148] to visualize EEG brain activity,[149] and to project computational Mueller matrix mapping results onto histological samples[150]. Similar to how contour lines enable data interpretation in geographical maps, we introduce the same concept to ion images to enable a multimodal IMS-microscopy data visualization strategy that integrates spatially, yet still yields easy interpretation by domain experts trained on only one of the two modalities.

In this work, we demonstrate the novel application of contour maps to depict ion intensity distributions as well as changes in ion intensity or derivatives of ion intensity to augment whole-slide histopathology images. We used soft-tissue *S. aureus* infection as a case study. A hallmark of *S. aureus* infection in tissue is the formation of abscesses. These abscesses can often look the identical under histopathological staining, yet still exhibit heterogeneous molecular signatures.[20,39,41,74] Combining molecular information from IMS measurements with the pathology information provided by PAS-stained microscopy can help elucidate changes in molecular architecture across the abscess that would otherwise remain difficult to discern.

RESULTS & DISCUSSION

A hallmark of soft-tissue staphylococcal infections is the formation of abscesses. Though these can be studied using histopathological staining such as PAS staining, approaches such as MALDI IMS can provide additional spatially resolved molecular information about the tissue. Elucidating the molecular changes in and around the staphylococcal abscess is critical to

understand how *S. aureus* proliferates and persists within host tissue, withstanding both host defenses and antibiotic drug treatments.

In this experiment, a mouse was inoculated with a fluorescently labeled strain of *S. aureus* and its kidney excised for analysis with PAS, MALDI IMS, and autofluorescence microscopy (Figure 4-1A). QuPath, a software tool for analyzing whole-slide images, was used to segment the fluorescently labeled *S. aureus*; in doing so, we identified one major abscess as well as smaller satellite infections throughout the kidney (Figure 4-1B). Regions of the kidney were annotated by a pathologist and include the major abscess, inflammatory cell infiltrate, renal pelvis, renal medulla, renal cortex, and adrenal gland (Figure 4-1C).

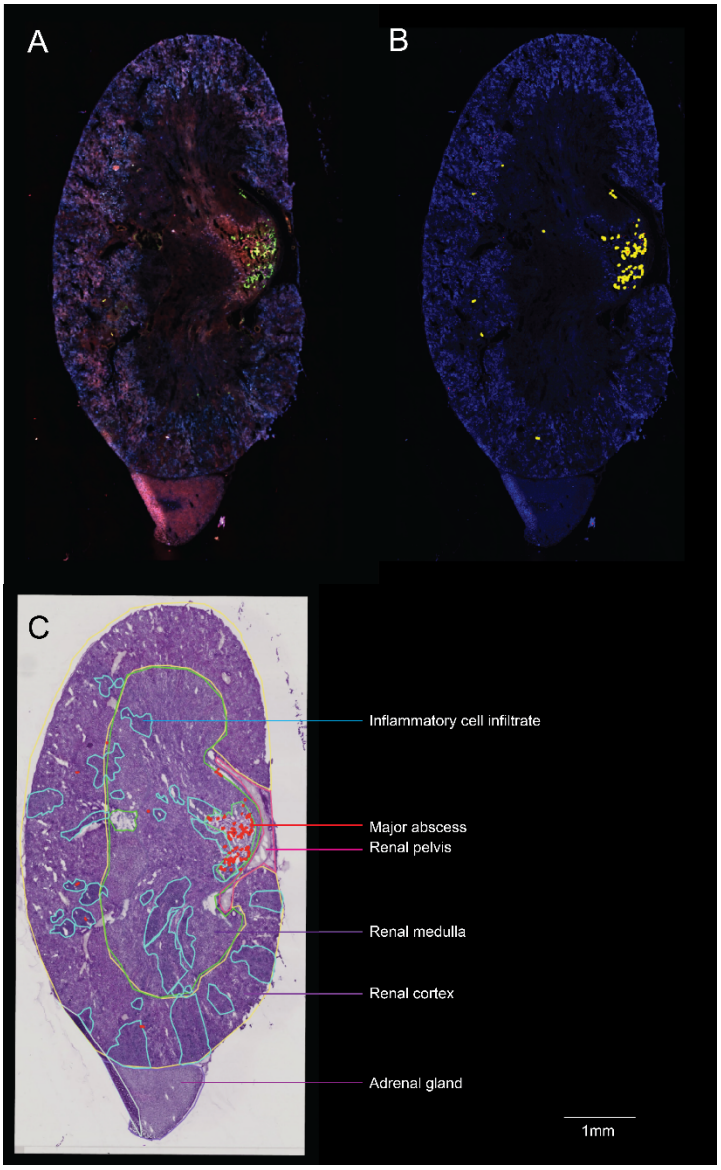

Figure 4-1: Murine kidney annotated by regions of interest. A) Autofluorescence microscopy image of a murine kidney; fluorescently labeled S. aureus can be seen by the green fluorescence. B) QuPath software was used to perform threshold-based segmentation of fluorescently labeled S. aureus, shown in yellow. C) Pathologist-annotated regions

of the kidney (renal pelvis, renal medulla, renal cortex, adrenal gland) and regions pertaining to infection (inflammatory cell infiltrate, major abscess).

Univariate Contour Maps

We used these labels to segment the IMS data and performed a pixel-wise Pearson correlation analysis among each metabolite from the IMS data and each major tissue structure. This univariate method allowed us to identify metabolites colocalizing to each of the annotated tissue structure types (Figure 4-2).

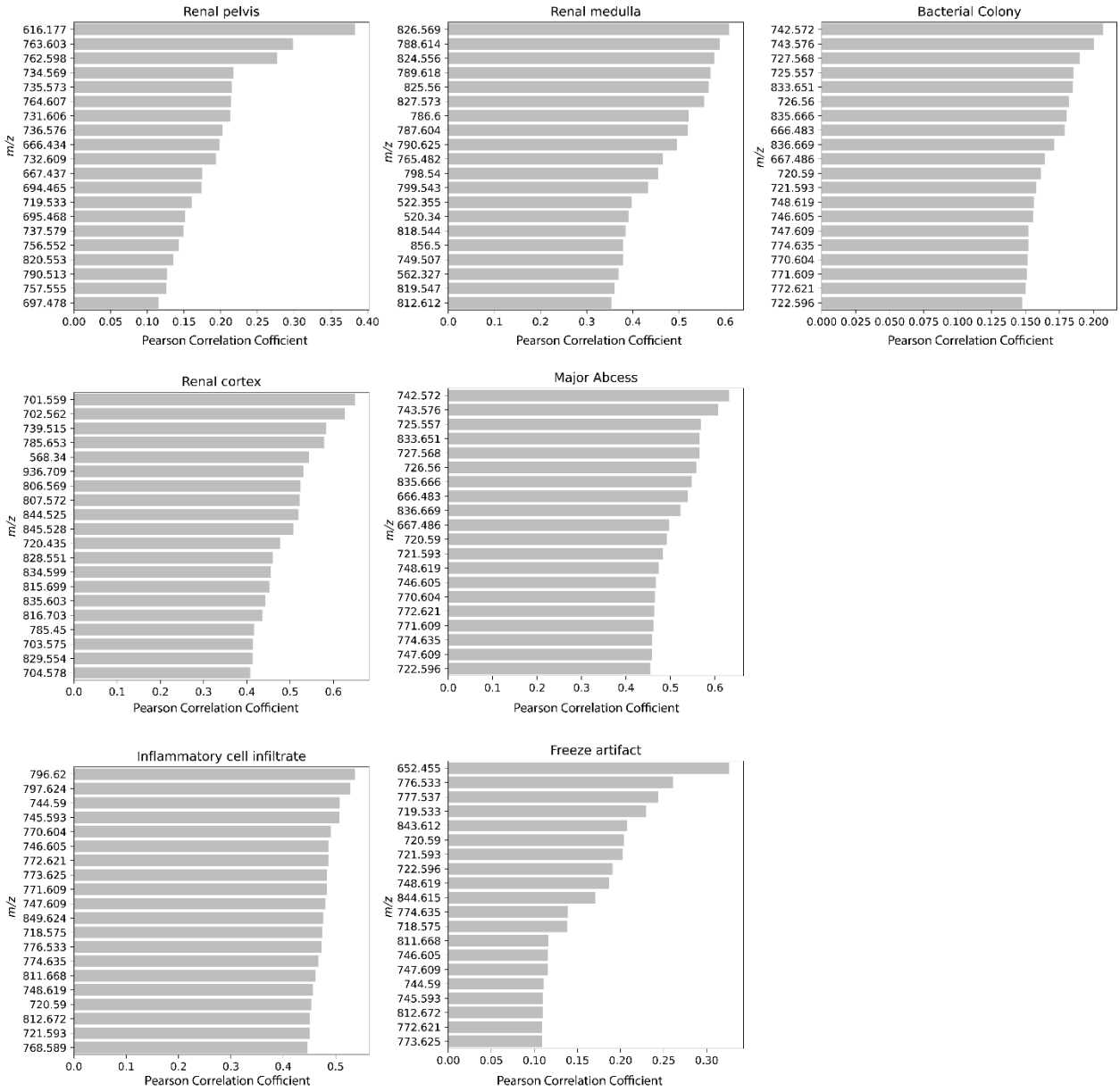

Figure 4-2: Pearson correlation coefficients for metabolites colocalizing with regions of the kidney and regions associated with infection.

For this portion of the work, we focused on the major abscess (Figure 4-3A). The Pearson correlation analysis revealed an ion (*m/z* 742.572, [PC(O-32:0)+Na]+, -0.09 ppm)) that colocalizes with bacterial colonies and the major abscess. The current approach to visualize the spatial localization of a single ion is to generate a heatmap of the ion intensity across the tissue (Figure 4-3B). To derive biological insight, stained microscopy images (e.g., PAS) and ion images are also commonly depicted next to each other. Alternatively, to facilitate easier cross-modal spatial mapping, the microscopy image is sometimes overlaid with the ion image into a single visualization with e.g. the opacity set to 50% for each. While the latter approach provides broad alignment and interpretation, overlaying colormaps of both modalities creates inherent ambiguities in terms of the dataset origin of an observed color in that blended visualization. For example, a pixel that is blue can be the result of a PAS violet-colored pixel mixing with an IMS green-colored pixel, or it could equally well be a mixing of a blue PAS pixel and a low-intensity and therefore nearly transparent IMS pixel. Opacity-based blended visualizations can be useful in certain scenarios, but when it is important to know the source of a particular shade of color or when specific intensity levels of the PAS or IMS data need to be discernable (because a domain expert is trained to use those levels or colors as cues for a recognition task), an opacity blending visualization is usually less well-suited. Using distinct colormaps for the different source modalities, such as one that is shades of one color and the other shades of a different color, is a possible solution; however, this approach still tends to lead to ambiguities for human observers since cognitive bandwidth would need to be used up to keep the color-origin mapping in mind while reading the image. Instead, we address this challenge by combining the two data sources into a single visualization, making spatial mapping between sources easier, but using different data visualization techniques to represent the content of each data source. Keeping the different content representations visually discernable, despite plotting on the same spatial coordinate system, allows for less ambiguous interpretation of the spatial connection between the distinct information sources despite populating the same visualization.

To do so, the normalized ion intensity data was binned, and contour lines were generated for each set of binned values (Figure 4-3C). These contour lines were then overlaid with the PAS (Figure 4-3D), providing an augmented visualization of the molecular changes of *m/z* 742.572 ([PC(O-32:0)+Na]+).across the major abscess.

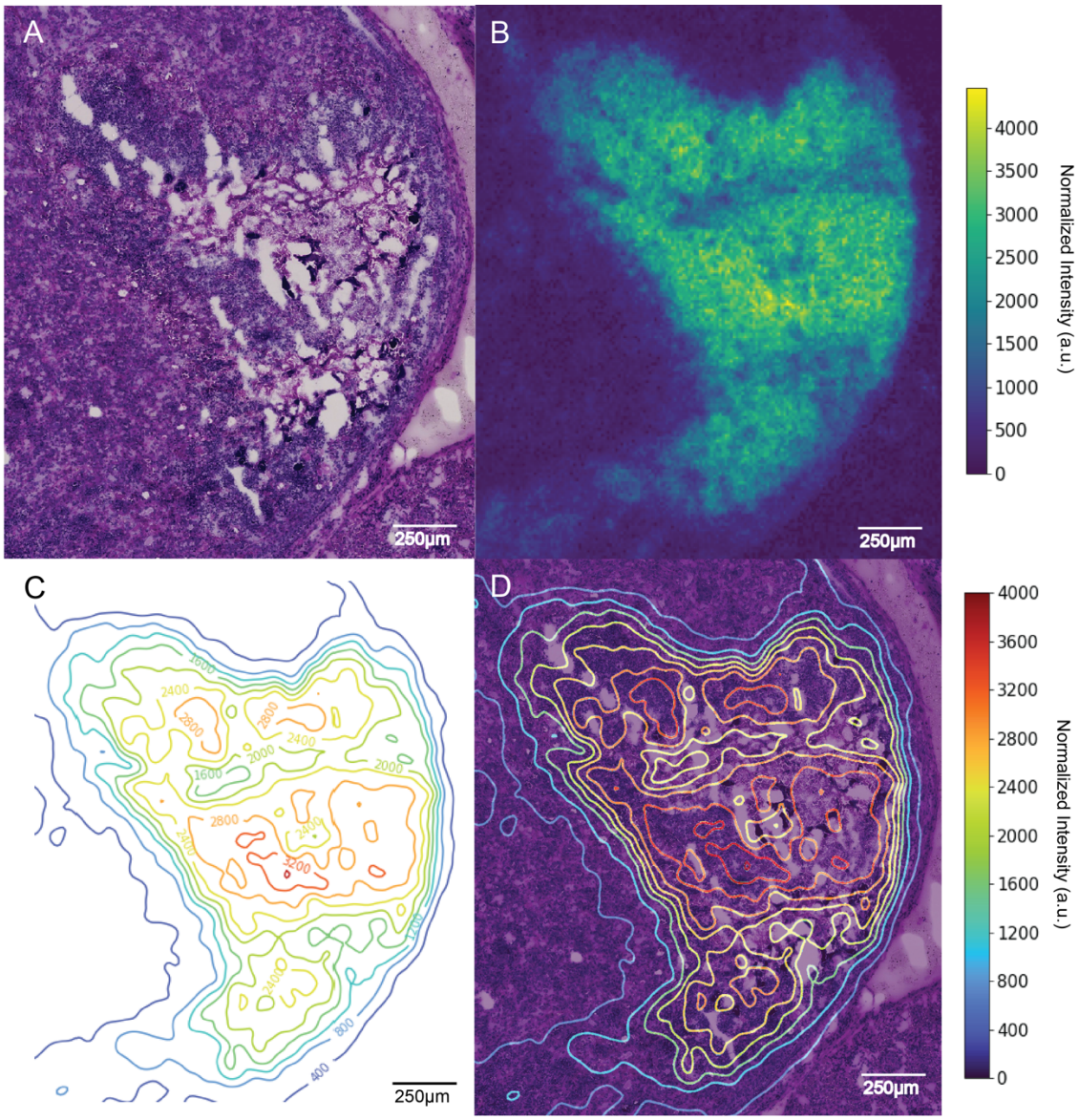

Figure 4-3: Contour map of a single MALDI IMS ion image correlating to a staphylococcal abscess. A) Periodic acid-Schiff (PAS) stain of an S. aureus abscess within a murine kidney. B) Heat map of an ion corresponding to lipid [PC(O-32:0)+Na]+ (m/z 742.572, -0.09 ppm) determined to co-localize with regions of staphylococcal infection based on a pixel-wise Pearson correlation analysis. C) Contour map generated using the ion image for PC(O-32:0), with contours labeled by binned m/z intensity values. D) Contour map overlaid with PAS.

Multivariate Contour Maps

Exploring IMS data in a univariate manner, focusing on a single *m/z* or molecular species at a time, is often useful in scenarios where there is prior knowledge on the molecular species that are relevant to the biological system at hand. However, in cases where such prior information is not available, separate consideration of each of the often hundreds of *m/z* distributions provided by an IMS experiment is often not practical. In those cases, it is usually more efficient to consider whether, among the hundreds of molecular species mapped, one can discern any multivariate trends that describe spatial and spectral variations that span multiple molecular species acting in

unison. Such a multivariate multi-species trend also has a spatial signature that describes where in the tissue that trend's panel of molecular species is heightened in intensity and where it is not. Similar to how the spatial distribution of a single ion (i.e., an ion image) can be represented as a contour plot, the spatial distribution of a multi-ion trend in the tissue (e.g., a component image, a cluster image, etc.) can be depicted as a contour plot. The latter will be referred to as a multivariate contour plot going forward.

Computational approaches allow us to take the hundreds of ion images and sort them using unsupervised methods such as principal component analysis (PCA), non-negative matrix factorization (NMF), *k*-means clustering, and spatial shrunken centroids clustering, thereby reducing the dimensionality of the IMS data and generating a subset of images that summarize molecular trends and spatial and/or spectral correlations within the measured data.[151–153] In this work, NMF was selected to go beyond single molecular species (or single *m/z)* images and to expand towards multivariate (or molecular panel) views into the molecular content of tissue. NMF is used to reveal a subset of underlying spatial and molecular trends or patterns in the IMS data, such that the empirically measured mass spectra and pixels are considered linear combinations of those underlying patterns. Such patterns or 'components' consist of a pseudo-spectrum, showing which molecular species along the *m/z* axis tend to appear in unison in the dataset, and a spatial distribution, which tells us where in the tissue those molecular species show that related ion intensity variation.

We performed NMF on the MALDI IMS data. Reconstruction errors were calculated for different hyperparameter settings, with the number of components *n* ranging from 1 to 100. The parameter value of $n=13$ components was determined to be optimal in this context, keeping the number of components low (enforcing a summarizing effect where a component bundles together several *m/z* variables) while also keeping the reconstruction error low (enforcing good approximation of the original measurements using the lower-dimensional NMF representation) (Figure 4-4).

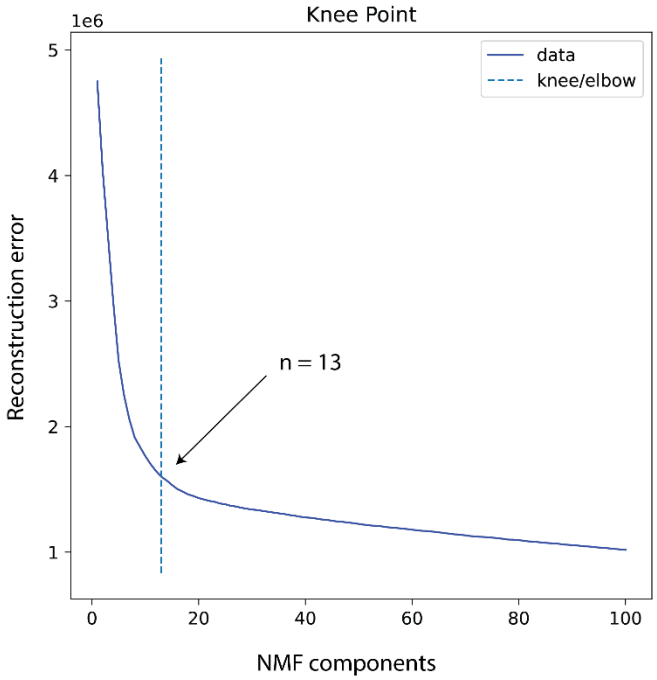

Figure 4-4: Reconstruction error of NMF components. Non-negative matrix factorization (NMF) was performed on the IMS data with components (n) ranging from 1 to 100. An n of 13 was determined to be optimal using the "elbow" method that indicates where the reconstruction error is minimized at the same time as the number of NMF components.

Upon performing NMF on this dataset allowing a rank-13 lower-dimensional approximation delivering underlying 13 components, we found that several of the components along the spatial axis correlated to infectious disease areas in the tissue, as well as to renal structures such as the medulla and cortex (Figure 4-5).

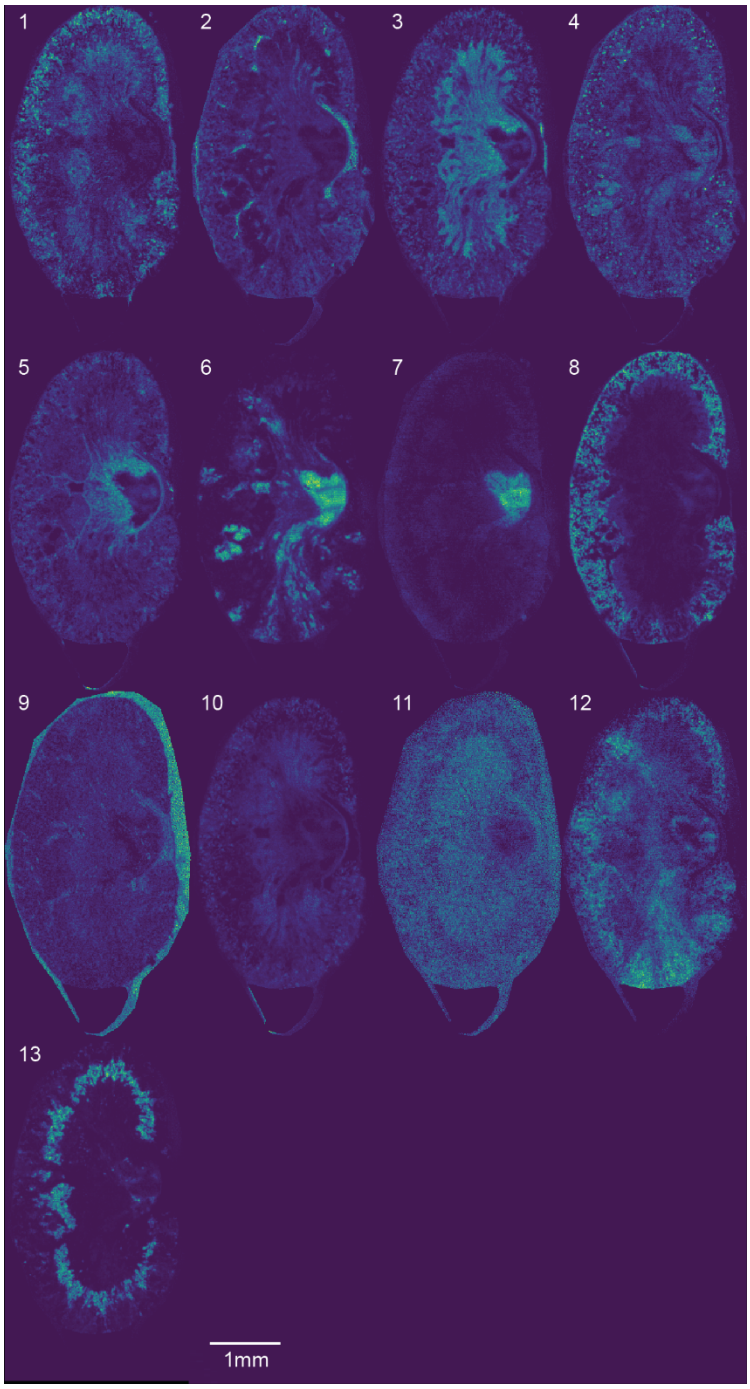

Figure 4-5: NMF components of IMS data. Non-negative matrix factorization (NMF) was performed on the IMS data and each of the resultant components were visualized as ion images.

In terms of infection, we found that NMF component 5 correlated with the inflammation region surrounding the abscess (Figure 4-6A,B), revealing key differences between the molecular composition of the inflammation and abscess. Conversely, NMF component 6 correlated with the major abscess (Figure 4-6C), revealing not only the spatially resolved molecular morphology of the abscess, but also the key areas of interest within the abscess based on the IMS data. For

instance, the top left region of the infection has a stronger molecular signal as compared to the rest of the abscess, as indicated by the red lines showing higher intensity, allowing us to speculate that this may be the center or origin point of the abscess. We can also observe two additional areas of high intensity, one towards the center of the abscess and one towards the bottom, further implying molecular heterogeneity in that region. Finally, studying these two components in the form of contour maps allows us to better visualize the transition zone at the host-pathogen interface. We can see that the transition zone is much narrower towards the top and right sides of the abscess than the left, evidenced by the increased proximity of contour lines towards the right, potentially indicating the direction in which the abscess was growing.

In summary, the contour maps for components correlating to the inflammation (Figure 4-6B) and abscess (Figure 4-6C), labeled with binned intensity values, reveal heterogeneous morphology in the tissue that was otherwise not discernable in the PAS alone. The average spectra corresponding to each component further reveals the distinct molecular profiles for the surrounding inflammation and bacterial abscess (Figure 4-6D).

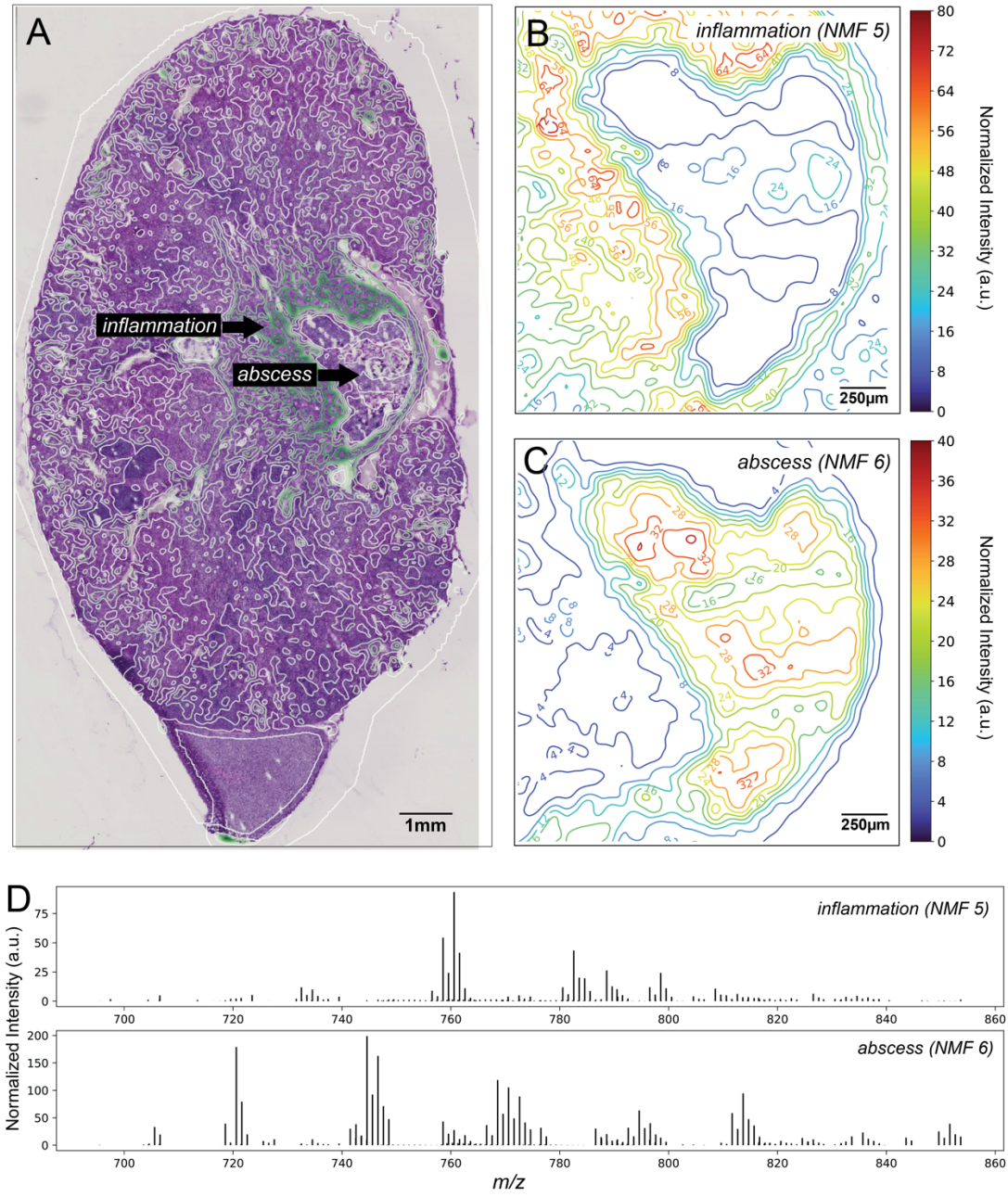

Figure 4-6: Contour maps built upon results of multivariate NMF analysis. Non-negative matrix factorization (NMF) was applied to the MALDI IMS data. A) Contour map of an NMF component (component 5) reveals distinct differentiation between the abscess and surrounding inflammation. B) Zoomed in contour map of an NMF component (component 5), which correlates to the surrounding inflammatory response; contours are labeled by binned intensity values. C) Another contour map of an NMF component (component 6) correlating to the bacterial abscess; contours are labeled by binned intensity values. D) Average spectra corresponding to each of the NMF components, revealing distinct molecular profiles between the surrounding inflammation (top) and bacterial abscess (bottom).

CONCLUSIONS

We have developed a new strategy to augment the visualization of digital pathology whole-slide images by co-registering contour maps of ion intensity or derivatives of ion intensity acquired by mass spectrometry. Within this work, contour maps were used to visualize spatially resolved molecular changes within an *S. aureus* infected murine kidney. Univariate contour maps built using single ion images revealed the spatial changes of individual ion intensity across whole slide images. Non-negative matrix factorization was applied to reduce the dimensionality of the MALDI IMS data, and multivariate contour maps of the NMF results were generated. These maps overlaid onto high-resolution stained microscopy exposed staphylococcal abscess morphology and the molecular transition zone at the host-pathogen interface, suggesting that one side of the abscess was more proliferative than the other. These findings were not discernable in the PAS nor IMS data alone, demonstrating the effectiveness of an augmented visualization of multimodal IMS data that provided insight into the molecular composition of an *S. aureus* infected murine kidney. Augmented visualization strategies such as this contour mapping approach are broadly applicable to multimodal IMS and microscopy experiments and will become more critical with continued advancement of multimodal workflows.

METHODS

Ethics statement

All animal experiments under protocol M1900043 were reviewed and approved by the Institutional Animal Care and Use Committee of Vanderbilt University. Procedures were performed according to the institutional policies, Animal Welfare Act, NIH guidelines, and American Veterinary Medical Association guidelines on euthanasia.

Murine model of systemic S. aureus infection

Six to eight-week-old female mice were anesthetized with tribromoethanol (Avertin) and retro-orbitally infected with ~1.5-2 x 107 CFUs of *S. aureus* USA300 LAC constitutively expressing sfGFP from the genome (PsarA-sfGFP integrated at the SaPI1 site). Infection was allowed to progress for seven days and then the animal was humanely euthanized, and the kidney was removed for molecular studies.

MALDI imaging mass spectrometry sample preparation and data acquisition

A 10 µm cryo-section of the infected murine kidney was mounted on a conductive ITO-coated slide, and 1,5-Diaminonapthalene was sprayed using an automated pneumatic TM sprayer from HTX Technologies.[154] Images were acquired in positive ionization mode at 10 µm raster size using a Bruker Daltonics timsTOF with beam-scanning mode turned off.[155] After MALDI IMS was acquired, the tissue section was scanned using autofluorescence (AF) microscopy with the matrix still present on the tissue section to reveal laser ablation marks of the MALDI IMS measurements to drive the IMS–microscopy image registration.[156]

Microscopy data acquisition

AF microscopy was performed before and after MALDI IMS. The AF images were captured using DAPI (ex. 335-383, em. 420-470), GFP (ex.450-490, em. 500-550), and DsRed (ex. 538-562, em. 570-640) filters on a Zeiss AxioScan.Z1 equipped with a Zeiss HXP 120V fluorescent metal-halide lamp. A 10x Plan-Apochromat (NA=0.45) objective was used, resulting

in a pixel side size of 0.65 μm/px when combined with a Hamamatsu ORCA flash monochromatic camera. Exposure times were 180, 142, and 242 milliseconds for the DAPI, GFP, and DsRed filters, respectively. Following this image acquisition, the slide was stained using a Periodic Acid-Schiff (PAS) stain according to existing protocols for human kidney tissue.[157] PAS-stained tissue sections were scanned using the Zeiss AxioScan.Z1 slide scanner using a 20x Plan-Aprochomat 20x (NA=0.8) objective with a Hitachi HV-F202SCL RGB camera for an effective pixel side size of 0.22 μm/px.

Microscopy-MALDI IMS image registration

Microscopy and MALDI IMS data were registered in two steps. In the first step, laser ablation marks captured by the post-IMS AF image were registered to MALDI IMS pixels using IMS MicroLink[158] by manually selecting corresponding pairs of laser ablation marks and IMS pixels. After directly mapping pixels to laser ablation marks in the microscopy pixel space, PAS and pre-MALDI AF microscopy images were registered to the IMS data via the previously registered post-IMS AF image in the wsireg[159] software.

Data Analysis

The registered PAS whole slide image (WSI) was imported into QuPath[34] and the kidney and infection regions were annotated by a domain expert. The remainder of the analysis was performed in Python version 3.7 and napari[35]. Pathology annotations were exported from QuPath using the GeoJSON standard and imported into Python, along with IMS data (TIC normalized using a 5-95% normalization). Lipid identifications were determined on the basis of high mass accuracy and matching to LIPIDMAPS lipidomics gateway (lipidmaps.org).[138,160] The adrenal gland was excluded from the analysis because the molecular species within this organ were vastly different from that of the kidney, with the potential of introducing extensive variation that could skew the analysis.

A pixel-wise Pearson correlation analysis was performed on the normalized intensity data and pathology annotations to distinguish ions highly correlating with each region of interest and to generate corresponding ion images. A Gaussian blur from the Sci-kit image[161] library was applied (σ = 2.0, truncate = 3.5) to the ion images to attenuate single-pixel variations and to emphasize multi-pixel variational patterns. Contour maps were generated from these smoothed images using the Matplotlib[162] library (levels = 10, all other hyper-parameters were set to the default). Non-negative matrix factorization (NMF) using Sci-kit learn's NMF implementation was applied to the full IMS dataset. To determine the optimal *n* number of components, a range of *n* values from 1 to 100 were tested using reconstruction error as a performance metric. The reconstruction error within this implementation is the Frobenius norm of the matrix difference between the training data and reconstructed data from the fitted model. Upon determining the optimal *n* value to be 13, where the number of components and reconstruction error were both minimized (Figure S4), the NMF algorithm was deployed onto the full IMS dataset and a result with 13 components was generated. A similar Gaussian blur (σ = 2.0, truncate = 3.5) and contour map method (levels = 10) was applied to the spatial signatures of these NMF components, generating contour maps. Output images were exported in the pyramidal OME-TIFF format for visualization in napari[35] alongside MALDI IMS and contour maps.

CHAPTER 5 CONCLUSIONS AND PERSPECTIVES

OVERVIEW

As imaging-based technology advances, the quantity of molecular data produced from a given tissue sample has increased immensely. Often, different molecular imaging modalities can be applied to the same or serial tissue section, generating orthogonal datasets that require computational methods to integrate, analyze, and visualize the molecular measurements. The research presented in this thesis establishes a series of technical developments in the field of biocomputational method development for integrated multimodal imaging (**Chapter 1**). First, an automated unsupervised method was established to analyze high-dimensional spatially targeted proteomic data utilizing PCA followed by *k*-means clustering while accounting for data sparsity (**Chapter 2**). Second, segmentation methods applied to CODEX MxIF data were evaluated, after which a custom multivariate unsupervised method was developed to segment multicellular FTUs and segmentation masks were used to extract FTU-specific lipidomic signatures from registered IMS data (**Chapter 3**). Third, a data visualization strategy was developed to augment classically stained microscopy images with IMS-based contour maps, providing an enhanced and more interpretable visualization of IMS-microscopy multimodal imaging datasets (**Chapter 4**).

These technical advances were applied to study *Staphylococcus aureus* infections within the murine kidney. The biological findings, comprising both proteomic and lipidomic molecular landscapes, provide new insight into the complex host-pathogen interface. First, the proteomic analysis from data obtained through microLESA revealed protein classes reflecting metabolic and cytoskeletal (**Chapter 2**). A biological hypothesis about the role of Annexins within the host-pathogen interface also emerged, suggesting that while Annexin 2 may be facilitating staphylococcal anchoring, Annexins 3 and 5 may confer varying degrees of protection against infection (**Chapter 2**). Second, analysis of multimodal CODEX MxIF and MALDI IMS data revealed lipidomic heterogeneity within staphylococcal abscesses as well as between regions with and without visible abscesses, with the identification of specific lipids localizing to each (**Chapter 3**). Third, a MALDI IMS-derived contour map visualization strategy overlaid onto PAS microscopy provided insight into the directionality of abscess growth and proliferation that was otherwise difficult to ascertain from each modality alone (**Chapter 4**). These results may inform the mechanism of staphylococcal abscess development and provide insight into the molecular pathways that are active and could potentially be disrupted to treat infection.

INSIGHT AND FUTURE STUDIES OF *S. AUREUS* INFECTION

The findings in this work allow for speculation about the host-pathogen interface of an *S. aureus* infection. The presence of infection is largely determined based on the interplay between the pathogen's proliferative action and resultant host immune response. If the pathogen is able to proliferate or persist at a faster rate than the host is able to eliminate the infection, the infection will endure, and disease severity may increase. Conversely, if the host can eliminate the infection faster than the pathogen can proliferate, the infection will be cleared. Both the host and pathogen have myriad mechanisms for clearing and persisting, respectively, and given the heterogenous molecular architecture of an *S. aureus* infection in tissue, these may be occurring differently across regions of the tissue.

The conclusions from the proteomics study suggest that Annexins play a key role in this interplay. Annexins are broadly understood as calcium-dependent phospholipid binding proteins and as a protein class, comprise twelve different proteins with varying functions. A survey of the literature suggested that Annexin A2 interacts with staphylococcal clumping factors A and B, facilitating attachment to epithelial cells. This physical anchoring of the bacteria through the binding of Annexin A2 may contribute to its long-lasting infection. The proteomics study detailed in this thesis (**Chapter 2**) revealed that Annexin A2 was increased in the interface and SAC 4 and 10 DPI. However, in addition to Annexin A2, Annexin A3 was also increased, suggesting it may also have a role in the host-pathogen interface. Little is known about the role of Annexin A3 within staphylococcal infection specifically; however, a transcriptomics study of blood samples from patients with sepsis indicated that Annexin A3 expression is restricted to neutrophils and is increased in septic samples compared to normal blood samples. This suggests that while Annexin A2 may be helping the bacteria remain in the tissue through an anchoring mechanism, Annexin A3 from the surrounding neutrophils could be providing a countereffect against the bacteria. In our proteomics study, it was also revealed that Annexin A5 was increased in only the interface and SAC samples 10 DPI. Like Annexin A3, Annexin A5 has not been implicated in staphylococcal infection but has been shown in previous studies to aid in the survival of a murine sepsis model. Specifically, Annexin A5 was found to inhibit HMGB1-mediated proinflammation and coagulation. Taken together, we can now speculate that Annexin A2 may be facilitating staphylococcal anchoring in the tissue while Annexins A3 and A5 provide varying degrees of protection. Future studies investigating these three Annexins could elucidate more clearly the impact each is having on the infection and reveal potential drug targets to bolster the host immune response and combat the infection.

The lipidomic studies detailed in this work largely focused on integrative methods to contextualize spatial lipid distributions with biologically relevant ROIs derived from microscopy (**Chapter 3, Chapter 4)**. Although we were able to make some preliminary identifications, including that of a phosphatidylcholine that co-localized with the bacterial abscess, there remains much to be discovered about lipids within the host-pathogen interface. Lipids form critical components of the lipid bilayer of both host and immune cells. Further, lipids have been implicated in antibiotic resistance of *S. aureus*. Within the work described here, we found that a subset of lipids such as $[SM(d34:1)+H]^+$ and $[PC(P\text{-}34:0)+H]^+$ co-localized with the abscess-rich regions while other lipids such as $[PC(36:1)+H]^+$ and $[PC(34:1)+H]^+$ co-localized with non-abscessed regions (**Chapter 3**). A broader characterization of all of the lipids that co-localize with different regions of infection can help elucidate molecular mechanisms of action at the host-pathogen interface. This characterization may include subsetting the full lipid dataset by lipid classes, chain length, or type of adduct and comparing differences among regions of the infection and/or FTUs and cell types. In the way that multiple molecular species were displayed as a single NMF component contour map (**Chapter 4**), these lipid distributions can then be overlaid onto PAS showing the histology and provide context about the localizations of different types of lipids. This information combined with the histological information indicating regions of infection could provide a more nuanced depiction of the lipidomic landscape at the host-pathogen interface.

FUTURE PERSPECTIVES

This work describes novel advances in biocomputational method development to integrate and analyze large spatially targeted multimodal imaging datasets using multivariate methods. However, technical advances are required before these methods can be applied to larger numbers of datasets in an automated way. These technical methods further require thoughtful considerations such as the selection of optimal hyper-parameters. Furthermore, the utility of these biocomputational approaches is to provide biological insight into spatially targeted molecular measurements by using one modality (often microscopy) to contextualize data from another (such as MALDI IMS). As such, there remains myriad opportunities to integrate additional types of spatially resolved data, such as that from spatial transcriptomics. Integrating transcriptomics data with proteomic and lipidomic data can further help elucidate biological processes, such as those at the host-pathogen interface and into other disease states.

The applications of the workflows described here allowed for a deeper analysis of the host-pathogen interface and provided novel insights into the complex biology that is often not captured in a single imaging modality. The biocomputational workflows described here would also be well-suited for data in other types of studies where there is a complex two-dimensional interplay of molecular species. One such example would be the study of the tumor microenvironment, where the heterogeneous molecular signatures can inform disease prognoses. These approaches could also be applied to study medical device implants, for instance to characterize molecular changes in the surrounding tissue as an IMS-derived contour map overlaid onto a histology image. A drug delivery device where gradients of delivered drugs and their pharmacokinetics and pharmacodynamics need to be monitored in a two- or three-dimensional space would also be a strong use-case for this type of multimodal workflow. Finally, these methods could be applied within a clinical setting. Often pathologists rely on tissue biopsies to rapidly diagnose diseases and delineate the region of infection. However, these assessments are typically based on histology alone. A future application of the methods detailed here could involve performing a MALDI IMS experiment on the same tissue section and building a contour map to visualize a known biomarker of disease. This assessment and subsequent visualization can inform the pathologist and/or surgeon resecting the diseased region more so than the histology alone can provide, resulting in a more complete resection and reducing the chances of disease resurgence.

REFERENCES

1. Fan, J., Han, F. & Liu, H. Challenges of Big Data analysis. *Natl. Sci. Rev.* **1**, 293–314 (2014).
2. Prentice, B. M., Caprioli, R. M. & Vuiblet, V. Label-free molecular imaging of the kidney. *Kidney Int.* **92**, 580–598 (2017).
3. Kruse, A. R. S. & Spraggins, J. M. Uncovering Molecular Heterogeneity in the Kidney With Spatially Targeted Mass Spectrometry. *Front. Physiol.* **13**, (2022).
4. Alturkistani, H. A., Tashkandi, F. M. & Mohammedsaleh, Z. M. Histological Stains: A Literature Review and Case Study. *Glob. J. Health Sci.* **8**, 72–79 (2015).
5. Javaeed, A. *et al.* Histological Stains in the Past, Present, and Future. *Cureus* **13**, e18486 (2021).
6. Jayapandian, C. P. *et al.* Development and evaluation of deep learning-based segmentation of histologic structures in the kidney cortex with multiple histologic stains. *Kidney Int.* **99**, 86–101 (2021).
7. Patterson, N. H. *et al.* Next Generation Histology-Directed Imaging Mass Spectrometry Driven by Autofluorescence Microscopy. *Anal. Chem.* **90**, 12404–12413 (2018).
8. Patterson, N. H., Tuck, M., Van De Plas, R. & Caprioli, R. M. Advanced Registration and Analysis of MALDI Imaging Mass Spectrometry Measurements through Autofluorescence Microscopy. *Anal. Chem.* **90**, 12395–12403 (2018).
9. Kim, S.-W., Roh, J. & Park, C.-S. Immunohistochemistry for Pathologists: Protocols, Pitfalls, and Tips. *J. Pathol. Transl. Med.* **50**, 411–418 (2016).
10. Goltsev, Y. *et al.* Deep Profiling of Mouse Splenic Architecture with CODEX Multiplexed Imaging In Brief A DNA barcoding-based imaging technique uses multiplexed tissue antigen staining to enable the characterization of cell types and dynamics in a model of autoimmune disease. *Cell* **174**, 968–981 (2018).
11. Schürch, C. M. *et al.* Coordinated Cellular Neighborhoods Orchestrate Antitumoral Immunity at the Colorectal Cancer Invasive Front. *Cell* **182**, 1341-1359.e19 (2020).
12. Kennedy-Darling, J. *et al.* Highly multiplexed tissue imaging using repeated oligonucleotide exchange reaction. *Eur. J. Immunol.* 1–32 (2021). doi:10.1002/eji.202048891
13. Pitt, J. J. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *Clin. Biochem. Rev.* **30**, 19–34 (2009).
14. Chen, G. & Pramanik, B. N. Application of LC/MS to proteomics studies: current status and future prospects. *Drug Discov. Today* **14**, 465–471 (2009).
15. Ryan, D. J. *et al.* MicroLESA: Integrating Autofluorescence Microscopy, in Situ Micro-Digestions, and Liquid Extraction Surface Analysis for High Spatial Resolution Targeted Proteomic Studies. *Anal. Chem.* **91**, 7578–7585 (2019).
16. Zhu, Y. *et al.* Spatially Resolved Proteome Mapping of Laser Capture Microdissected Tissue with Automated Sample Transfer to Nanodroplets*. *Mol. Cell. Proteomics* **17**, 1864–

1874 (2018).

17. Zhu, Y. *et al.* Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nat. Commun. 2018 91* **9**, 1–10 (2018).

18. Piehowski, P. D. *et al.* Automated mass spectrometry imaging of over 2000 proteins from tissue sections at 100-μm spatial resolution. *Nat. Commun.* **11**, 1–12 (2020).

19. Guiberson, E. R. *et al.* Spatially Targeted Proteomics of the Host–Pathogen Interface during Staphylococcal Abscess Formation. *ACS Infect. Dis.* **7**, 101–113 (2021).

20. Sharman, K. *et al.* Rapid Multivariate Analysis Approach to Explore Differential Spatial Protein Profiles in Tissue. *J. Proteome Res.* (2022). doi:10.1021/acs.jproteome.2c00206

21. Caprioli, R. M., Farmer, T. B. & Gile, J. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Anal. Chem.* **69**, 4751–4760 (1997).

22. Norris, J. L. & Caprioli, R. M. Analysis of tissue specimens by matrix-assisted laser desorption/ionization imaging mass spectrometry in biological and clinical research. *Chem. Rev.* **113**, 2309–2342 (2013).

23. Fujino, Y., Minamizaki, T., Yoshioka, H., Okada, M. & Yoshiko, Y. Imaging and mapping of mouse bone using MALDI-imaging mass spectrometry. *Bone reports* **5**, 280–285 (2016).

24. Judd, A. M. *et al.* A recommended and verified procedure for in situ tryptic digestion of formalin-fixed paraffin-embedded tissues for analysis by matrix-assisted laser desorption/ionization imaging mass spectrometry. *J. Mass Spectrom.* **54**, 716–727 (2019).

25. Dufresne, M., Patterson, N. H., Norris, J. L. & Caprioli, R. M. Combining Salt Doping and Matrix Sublimation for High Spatial Resolution MALDI Imaging Mass Spectrometry of Neutral Lipids. *Anal. Chem.* (2019). doi:10.1021/acs.analchem.9b02974

26. Angel, P. M. *et al.* Advances in MALDI imaging mass spectrometry of proteins in cardiac tissue, including the heart valve. *Biochim. Biophys. acta. Proteins proteomics* **1865**, 927–935 (2017).

27. Powers, T. W. *et al.* MALDI Imaging Mass Spectrometry Profiling of N-Glycans in Formalin-Fixed Paraffin Embedded Clinical Tissue Blocks and Tissue Microarrays. *PLoS One* **9**, e106255 (2014).

28. Zavalin, A., Yang, J., Hayden, K., Vestal, M. & Caprioli, R. M. Tissue protein imaging at 1 μm laser spot diameter for high spatial resolution and high imaging speed using transmission geometry MALDI TOF MS. *Anal. Bioanal. Chem.* **407**, 2337–2342 (2015).

29. Alexandrov, T. MALDI imaging mass spectrometry: statistical data analysis and current computational challenges. *BMC Bioinformatics* **13 Suppl 1**, S11 (2012).

30. Hu, H. & Laskin, J. Emerging Computational Methods in Mass Spectrometry Imaging. *Adv. Sci.* **n/a**, 2203339 (2022).

31. Van De Plas, R., Yang, J., Spraggins, J. & Caprioli, R. M. Fusion of mass spectrometry and microscopy: a multi-modality paradigm for molecular tissue mapping HHS Public Access. *Nat Methods* **12**, 366–372 (2015).

32. Verbeeck, N., Caprioli, R. M. & Van de Plas, R. Unsupervised machine learning for

exploratory data analysis in imaging mass spectrometry. *Mass Spectrom. Rev.* **39**, 245–291 (2020).

33. Tideman, L. E. M. *et al.* Automated biomarker candidate discovery in imaging mass spectrometry data through spatially localized Shapley additive explanations. *Anal. Chim. Acta* **1177**, 338522 (2021).

34. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 1–7 (2017).

35. Sofroniew, N. *et al.* Napari: A multi-dimensional image viewer for python. (2019). doi:10.5281/ZENODO.4048613

36. CDC. Deadly Staph Infections Still Threaten the U.S. CDC calls for increased prevention to protect patients. *Vital Signs* (2019). Available at: https://www.cdc.gov/media/releases/2019/p0305-deadly-staph-infections.html. (Accessed: 27th September 2020)

37. Pfizer Inc. New Research Estimates MRSA Infections Cost U.S. Hospitals $3.2 Billion to $4.2 Billion Annually. *Infection Control Today* 1–2 (2005). Available at: https://www.infectioncontroltoday.com/view/new-research-estimates-mrsa-infections-cost-us-hospitals-32-billion-42-billion-annually. (Accessed: 27th September 2020)

38. Casadevall, A. & Pirofski, L. A. Host-pathogen interactions: Basic concepts of microbial commensalism, colonization, infection, and disease. *Infection and Immunity* **68**, 6511–6518 (2000).

39. Cassat, J. E. *et al.* Integrated molecular imaging reveals tissue heterogeneity driving host-pathogen interactions. *Sci. Transl. Med.* **10**, 6361 (2018).

40. Cheng, A. G., DeDent, A. C., Schneewind, O. & Missiakas, D. A play in four acts: Staphylococcus aureus abscess formation. *Trends in Microbiology* **19**, 225–232 (2011).

41. Perry, W. J. *et al.* Staphylococcus aureus exhibits heterogeneous siderophore production within the vertebrate host. *Proc. Natl. Acad. Sci. U. S. A.* **116**, 21980–21982 (2019).

42. Surewaard, B. G. J. *et al.* Identification and treatment of the Staphylococcus aureus reservoir in vivo. *J. Exp. Med.* **213**, 1141–1151 (2016).

43. Aebersold, R. & Mann, M. Mass spectrometry-based proteomics. *Nature* **422**, 198–207 (2003).

44. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J. & Kuster, B. Quantitative mass spectrometry in proteomics: a critical review. *Anal. Bioanal. Chem.* **389**, 1017–1031 (2007).

45. Han, X., Aslanian, A. & Yates 3rd, J. R. Mass spectrometry for proteomics. *Curr. Opin. Chem. Biol.* **12**, 483–490 (2008).

46. Yates, J. R., Ruse, C. I. & Nakorchevsky, A. Proteomics by Mass Spectrometry: Approaches, Advances, and Applications. *Annu. Rev. Biomed. Eng.* **11**, 49–79 (2009).

47. Baker, E. S. *et al.* Mass spectrometry for translational proteomics: progress and clinical implications. *Genome Med.* **4**, 63 (2012).

48. Noor, Z., Ahn, S. B., Baker, M. S., Ranganathan, S. & Mohamedali, A. Mass spectrometry-based protein identification in proteomics- A review. *Briefings in Bioinformatics* **22**, 1620–

1638 (2021).

49. Srinivas, P. R., Verma, M., Zhao, Y. & Srivastava, S. Proteomics for Cancer Biomarker Discovery. *Clin. Chem.* **48**, 1160–1169 (2002).

50. Sallam, R. M. Proteomics in Cancer Biomarkers Discovery: Challenges and Applications. *Dis. Markers* **2015**, 321370 (2015).

51. Shruthi, B. S., Vinodhkumar, P. & Selvamani. Proteomics: A new perspective for cancer. *Adv. Biomed. Res.* **5**, 67 (2016).

52. Scott, E. M., Carter, A. M. & Findlay, J. B. C. The application of proteomics to diabetes. *Diabetes Vasc. Dis. Res.* **2**, 54–60 (2005).

53. Bhat, S., Jagadeeshaprasad, M. G., Venkatasubramani, V. & Kulkarni, M. J. Abundance matters: role of albumin in diabetes, a proteomics perspective. *Expert Rev. Proteomics* **14**, 677–689 (2017).

54. Wang, N., Zhu, F., Chen, L. & Chen, K. Proteomics, metabolomics and metagenomics for type 2 diabetes and its complications. *Life Sci.* **212**, 194–202 (2018).

55. Fu, J. *et al.* Advances in Current Diabetes Proteomics: From the Perspectives of Label- free Quantification and Biomarker Selection. *Curr. Drug Targets* **21**, 34–54 (2020).

56. Van Eyk, J. E. Proteomics: unraveling the complexity of heart disease and striving to change cardiology. *Curr. Opin. Mol. Ther.* **3**, 546–553 (2001).

57. McGregor, E. & Dunn, M. J. Proteomics of the heart: unraveling disease. *Circ. Res.* **98**, 309–321 (2006).

58. Fu, Q. & Van Eyk, J. E. Proteomics and heart disease: identifying biomarkers of clinical utility. *Expert Rev. Proteomics* **3**, 237–249 (2006).

59. Baetta, R., Pontremoli, M., Martinez Fernandez, A., Spickett, C. M. & Banfi, C. Proteomics in cardiovascular diseases: Unveiling sex and gender differences in the era of precision medicine. *J. Proteomics* **173**, 62–76 (2018).

60. Stoeckli, M., Chaurand, P., Hallahan, D. E. & Caprioli, R. M. Imaging mass spectrometry: a new technology for the analysis of protein expression in mammalian tissues. *Nat. Med.* **7**, 493–496 (2001).

61. Amstalden van Hove, E. R., Smith, D. F. & Heeren, R. M. A. A concise review of mass spectrometry imaging. *J. Chromatogr. A* **1217**, 3946–3954 (2010).

62. McDonnell, L. A. & Heeren, R. M. A. Imaging mass spectrometry. *Mass Spectrom. Rev.* **26**, 606–643 (2007).

63. Burnum, K. E., Frappier, S. L. & Caprioli, R. M. Matrix-assisted laser desorption/ionization imaging mass spectrometry for the investigation of proteins and peptides. *Annu. Rev. Anal. Chem. (Palo Alto. Calif).* **1**, 689–705 (2008).

64. Spraggins, J. M. *et al.* Next-generation technologies for spatial proteomics: Integrating ultra-high speed MALDI-TOF and high mass resolution MALDI FTICR imaging mass spectrometry for protein analysis. *Proteomics* **16**, 1678–1689 (2016).

65. Aichler, M. & Walch, A. MALDI Imaging mass spectrometry: current frontiers and

perspectives in pathology research and practice. *Lab. Investig.* **95**, 422–431 (2015).

66. Ryan, D. J., Spraggins, J. M. & Caprioli, R. M. Protein identification strategies in MALDI imaging mass spectrometry: a brief review. *Curr. Opin. Chem. Biol.* **48**, 64–72 (2019).

67. Kelly, R. *et al.* Single Cell Proteome Mapping of Tissue Heterogeneity Using Microfluidic Nanodroplet Sample Processing and Ultrasensitive LC-MS. *J. Biomol. Tech.* **30**, S61 (2019).

68. Williams, S. M. *et al.* Automated Coupling of Nanodroplet Sample Preparation with Liquid Chromatography-Mass Spectrometry for High-Throughput Single-Cell Proteomics. *Anal. Chem.* **92**, 10588–10596 (2020).

69. Sarsby, J. *et al.* Liquid Extraction Surface Analysis Mass Spectrometry Coupled with Field Asymmetric Waveform Ion Mobility Spectrometry for Analysis of Intact Proteins from Biological Substrates. *Anal. Chem.* **87**, 6794–6800 (2015).

70. Schey, K. L., Anderson, D. M. & Rose, K. L. Spatially-directed protein identification from tissue sections by top-down LC-MS/MS with electron transfer dissociation. *Anal. Chem.* **85**, 6767–6774 (2013).

71. Wisztorski, M. *et al.* Droplet-based liquid extraction for spatially-resolved microproteomics analysis of tissue sections. in *Methods in Molecular Biology* **1618**, 49–63 (Humana Press Inc., 2017).

72. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).

73. Tyanova, S., Temu, T. & Cox, J. The MaxQuant computational platform for mass spectrometry-based shotgun proteomics. *Nat. Protoc.* **11**, 2301–2319 (2016).

74. Guiberson, E. R. *et al.* Spatially Targeted Proteomics of the Host-Pathogen Interface during Staphylococcal Abscess Formation. *ACS Infect. Dis.* **7**, 101–113 (2021).

75. Piehowski, P. D. *et al.* Automated mass spectrometry imaging of over 2000 proteins from tissue sections at 100-μm spatial resolution. *Nat. Commun.* **11**, 8 (2020).

76. Satoskar, A. A. *et al.* Characterization of Glomerular Diseases Using Proteomic Analysis of Laser Capture Microdissected Glomeruli. *Mod Pathol* **25**, 709–721 (2012).

77. Cazares, L. H. *et al.* Normal, benign, preneoplastic, and malignant prostate cells have distinct protein expression profiles resolved by Surface Enhanced Laser Desorption/Ionization mass spectrometry. *Clin. Cancer Res.* **8**, 2541–2552 (2002).

78. Datta, S. *et al.* Laser capture microdissection: Big data from small samples. *Histol. Histopathol.* **30**, 1255–1269 (2015).

79. Schuetz, C. S. *et al.* Progression-specific genes identified by expression profiling of matched ductal carcinomas in situ and invasive breast tumors, combining laser capture microdissection and oligonucleotide microarray analysis. *Cancer Res.* **66**, 5278–5286 (2006).

80. Alevizos, I. *et al.* Oral cancer in vivo gene expression profiling assisted by laser capture microdissection and microarray analysis. *Oncogene* **20**, 6196–6204 (2001).

81. Kunz, G. M. & Chan, D. W. The use of laser capture microscopy in proteomics research - A review. *Dis. Markers* **20**, 155–160 (2004).

82. Shapiro, J. P. *et al.* A quantitative proteomic workflow for characterization of frozen clinical biopsies: Laser capture microdissection coupled with label-free mass spectrometry. *J. Proteomics* **77**, 433–440 (2012).

83. Elias, J. *et al.* Prevalence dependent calibration of a predictive model for nasal carriage of methicillin-resistant Staphylococcus aureus. *BMC Infect. Dis.* **13**, 111 (2013).

84. Wei, R. *et al.* Missing Value Imputation Approach for Mass Spectrometry-based Metabolomics Data. *Sci. Rep.* **8**, 663 (2018).

85. Dabke, K., Kreimer, S., Jones, M. R. & Parker, S. J. A simple optimization workflow to enable precise and accurate imputation of missing values in proteomic datasets. *J. Proteome Res.* **20**, 3214–3229 (2021).

86. Lazar, C., Gatto, L., Ferro, M., Bruley, C. & Burger, T. Accounting for the Multiple Natures of Missing Values in Label-Free Quantitative Proteomics Data Sets to Compare Imputation Strategies. *J. Proteome Res.* **15**, 1116–1125 (2016).

87. Anderson, D. C., Li, W., Payan, D. G. & Noble, W. S. A new algorithm for the evaluation of shotgun peptide sequencing in proteomics: Support vector machine classification of peptide MS/MS spectra and SEQUEST scores. *J. Proteome Res.* **2**, 137–146 (2003).

88. Klein, O. *et al.* MALDI-Imaging for Classification of Epithelial Ovarian Cancer Histotypes from a Tissue Microarray Using Machine Learning Methods. *Proteomics - Clin. Appl.* **13**, 1–11 (2019).

89. Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S. & Bacardit, J. Application of Machine Learning to Proteomics Data: Classification and Biomarker Identification in Postgenomics Biology. *Omi. A J. Integr. Biol.* **17**, 595–610 (2013).

90. Halko, N., Martinsson, P. G. & Tropp, J. A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM Rev.* **53**, 217–288 (2011).

91. Hastie, T., Tibshirani, R. & Friedman, J. H. *The elements of statistical learning data mining, inference, and prediction*. *Springer series in statistics,* (Springer, 2009).

92. Sinitcyn, P., Rudolph, J. D. & Cox, J. Computational Methods for Understanding Mass Spectrometry–Based Shotgun Proteomics Data. *Annu. Rev. Biomed. Data Sci.* **1**, 207–234 (2018).

93. Hartigan, J. A. & Wong, M. A. Algorithm AS 136: A K-Means Clustering Algorithm. *Appl. Stat.* **28**, 100 (1979).

94. Lloyd, S. P. Least Squares Quantization in PCM. *IEEE Trans. Inf. Theory* **28**, 129–137 (1982).

95. MacQueen, J. Some methods for classification and analysis of multivariate observations. in *Proceedings of the fifth Berkeley Symposium on Mathematical Statistics and Probability* **1**, 281–296 (The Regents of the University of California, 1967).

96. Rousseeuw, P. J. Silhouettes: A graphical aid to the interpretation and validation of cluster

analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).

97. Karpievitch, Y. *et al.* A statistical framework for protein quantitation in bottom-up MS-based proteomics. *Bioinformatics* **25**, 2028–2034 (2009).

98. Santamaria-Kisiel, L., Rintala-Dempsey, A. C. & Shaw, G. S. Calcium-dependent and -independent interactions of the S100 protein family. *Biochem. J.* **396**, 201–214 (2006).

99. Mi, H. *et al.* PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **49**, D394–D403 (2021).

100. Ziesemer, S. *et al.* Staphylococcus aureus α-Toxin Induces Actin Filament Remodeling in Human Airway Epithelial Model Cells. *Am. J. Respir. Cell Mol. Biol.* **58**, 482–491 (2018).

101. Gotoh, M. *et al.* Annexins I and IV inhibit Staphylococcus aureus attachment to human macrophages. *Immunol. Lett.* **98**, 297–302 (2005).

102. Ashraf, S., Cheng, J. & Zhao, X. Clumping factor A of Staphylococcus aureus interacts with AnnexinA2 on mammary epithelial cells. *Sci. Rep.* **7**, 40608 (2017).

103. Ying, Y.-T. *et al.* Annexin A2-Mediated Internalization of Staphylococcus aureus into Bovine Mammary Epithelial Cells Requires Its Interaction with Clumping Factor B. *Microorganisms* **9**, 2090 (2021).

104. He, X. *et al.* A new role for host annexin A2 in establishing bacterial adhesion to vascular endothelial cells: lines of evidence from atomic force microscopy and an in vivo study. *Lab. Investig.* **99**, 1650–1660 (2019).

105. Toufiq, M. *et al.* Annexin A3 in sepsis: novel perspectives from an exploration of public transcriptome data. *Immunology* **161**, 291–302 (2020).

106. Park, J. H. *et al.* Annexin A5 increases survival in murine sepsis model by inhibiting HMGB1-mediated pro-inflammation and coagulation. *Mol. Med.* **22**, 424–436 (2016).

107. Randall, E. C., Race, A. M., Cooper, H. J. & Bunch, J. MALDI Imaging of Liquid Extraction Surface Analysis Sampled Tissue. *Anal. Chem.* **88**, 8433–8440 (2016).

108. Kertesz, V. & Van Berkel, G. J. Liquid microjunction surface sampling coupled with high-pressure liquid chromatography-electrospray ionization-mass spectrometry for analysis of drugs and metabolites in whole-body thin tissue sections. *Anal. Chem.* **82**, 5917–5921 (2010).

109. Kertesz, V., Weiskittel, T. M. & Van Berkel, G. J. An enhanced droplet-based liquid microjunction surface sampling system coupled with HPLC-ESI-MS/MS for spatially resolved analysis. *Anal. Bioanal. Chem.* **407**, 2117–2125 (2015).

110. Parkinson, E. *et al.* Proteomic analysis of the human skin proteome after In Vivo treatment with sodium dodecyl sulphate. *PLoS One* **9**, e97772 (2014).

111. Bliss, E., Heywood, W. E., Benatti, M., Sebire, N. J. & Mills, K. An optimised method for the proteomic profiling of full thickness human skin. *Biol. Proced. Online* **18**, 15 (2016).

112. Simone, N. L. *et al.* Sensitive immunoassay of tissue cell proteins procured by laser capture microdissection. *Am. J. Pathol.* **156**, 445–452 (2000).

113. Harris, G. A., Nicklay, J. J. & Caprioli, R. M. Localized in situ hydrogel-mediated protein

digestion and extraction technique for on-tissue analysis. *Anal. Chem.* **85**, 2717–2723 (2013).

114. Taverna, D., Norris, J. L. & Caprioli, R. M. Histology-directed microwave assisted enzymatic protein digestion for MALDI ms analysis of mammalian tissue. *Anal. Chem.* **87**, 670–676 (2015).

115. Nicklay, J. J., Harris, G. A., Schey, K. L. & Caprioli, R. M. MALDI imaging and in situ identification of integral membrane proteins from rat brain tissue sections. *Anal. Chem.* **85**, 7191–7196 (2013).

116. The UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).

117. Patterson, N. H., Tuck, M., Van De Plas, R. & Caprioli, R. M. Advanced Registration and Analysis of MALDI Imaging Mass Spectrometry Measurements through Autofluorescence Microscopy. *Anal. Chem.* **90**, 12395–12403 (2018).

118. Kachouie, N. N., Fieguth, P. & Jervis, E. Watershed deconvolution for cell segmentation. *Proc. 30th Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. EMBS'08 - "Personalized Healthc. through Technol.* 375–378 (2008). doi:10.1109/iembs.2008.4649168

119. Carpenter, A. E. *et al.* CellProfiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biol.* **7**, R100 (2006).

120. McQuin, C. *et al.* CellProfiler 3.0: Next-generation image processing for biology. *PLOS Biol.* **16**, e2005970 (2018).

121. Czech, E., Aksoy, A., Aksoy, P. & Hammerbacher, J. Cytokit: a single-cell analysis toolkit for high dimensional fluorescent microscopy imaging. doi:10.1186/s12859-019-3055-3

122. Sommer, C., Straehle, C., Köthe, U. & Hamprecht, F. A. Ilastik: Interactive learning and segmentation toolkit. in *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro* 230–233 (2011). doi:10.1109/ISBI.2011.5872394

123. Berg, S. *et al.* Ilastik: Interactive Machine Learning for (Bio)Image Analysis. *Nat. Methods* **16**, 1226–1232 (2019).

124. Schapiro, D. *et al.* histoCAT: analysis of cell phenotypes and interactions in multiplex image cytometry data. *Nat. Methods* **14**, 873–876 (2017).

125. Guiberson, E. R. *et al.* Spatially-targeted proteomics of the host-pathogen interface during staphylococcal abscess 11 12. *bioRxiv* 2020.09.01.267773 (2020). doi:10.1101/2020.09.01.267773

126. Hubler, M. J. & Kennedy, A. J. Role of lipids in the metabolism and activation of immune cells. *J. Nutr. Biochem.* **34**, 1–7 (2016).

127. Varshney, P., Yadav, V. & Saini, N. Lipid rafts in immune signalling: current progress and future perspective. *Immunology* **149**, 13–24 (2016).

128. Chiurchiù, V. *et al.* Proresolving lipid mediators resolvin D1, resolvin D2, and maresin 1 are critical in modulating T cell responses. *Sci. Transl. Med.* **8**, 353ra111 (2016).

129. Pan, Y. *et al.* Survival of tissue-resident memory T cells requires exogenous lipid uptake and metabolism. *Nature* **543**, 252–256 (2017).

130. Braverman, N. E. & Moser, A. B. Functions of plasmalogen lipids in health and disease. *Biochim. Biophys. Acta* **1822**, 1442–1452 (2012).

131. Record, M., Silvente-Poirot, S., Poirot, M. & Wakelam, M. J. O. Extracellular vesicles: lipids as key components of their biogenesis and functions. *J. Lipid Res.* **59**, 1316–1324 (2018).

132. Bisignano, C. *et al.* Study of the Lipid Profile of ATCC and Clinical Strains of Staphylococcus aureus in Relation to Their Antibiotic Resistance. *Molecules* **24**, (2019).

133. Rocklin, M. Dask: Parallel Computation with Blocked algorithms and Task Scheduling. in *Proceedings of the 14th Python in Science Conference* 126–132 (2015). doi:10.25080/majora-7b98e3ed-013

134. Sculley, D. Web-Scale K-Means Clustering.

135. Munro, B. Manual of Histologic Staining Methods of the Armed Forces Institute of Pathology. *Pathology* **3**, 249 (1971).

136. Sud, M. *et al.* LMSD: LIPID MAPS structure database. *Nucleic Acids Res.* **35**, D527–D532 (2007).

137. Fahy, E. *et al.* Update of the LIPID MAPS comprehensive classification system for lipids1. *J. Lipid Res.* **50**, S9–S14 (2009).

138. Fahy, E., Sud, M., Cotter, D. & Subramaniam, S. LIPID MAPS online tools for lipid research. *Nucleic Acids Res.* **35**, W606–W612 (2007).

139. Neumann, E. K. *et al.* Highly multiplexed immunofluorescence of the human kidney using co-detection by indexing. *Kidney Int.* (2021). doi:https://doi.org/10.1016/j.kint.2021.08.033

140. Miles, A. *et al.* zarr-developers/zarr-python: v2.13.3. (2022). doi:10.5281/ZENODO.7174882

141. Caprioli, R. M., Farmer, T. B. & Gile, J. Molecular Imaging of Biological Samples: Localization of Peptides and Proteins Using MALDI-TOF MS. *Anal. Chem.* (1997). doi:10.1021/ac970888i

142. Spraggins, J. M. *et al.* MALDI FTICR IMS of Intact Proteins: Using Mass Accuracy to Link Protein Images with Proteomics Data. *J. Am. Soc. Mass Spectrom.* **26**, 974–985 (2015).

143. Longuespée, R. *et al.* MALDI mass spectrometry imaging: A cutting-edge tool for fundamental and clinical histopathology. *PROTEOMICS – Clin. Appl.* **10**, 701–719 (2016).

144. Nilsson, A. *et al.* Mass spectrometry imaging in drug development. *Anal. Chem.* **87**, 1437–1455 (2015).

145. Heeren, R. M. A. Getting the picture: The coming of age of imaging MS. *Int. J. Mass Spectrom.* **377**, 672–680 (2015).

146. Patterson, N. H. *et al.* Autofluorescence microscopy as a label-free tool for renal histology and glomerular segmentation. *bioRxiv* 2021.07.16.452703 (2021). doi:10.1101/2021.07.16.452703

147. Van de Plas, R., Yang, J., Spraggins, J. & Caprioli, R. M. Image fusion of mass spectrometry and microscopy: a multimodality paradigm for molecular tissue mapping. *Nat. Methods* **12**,

366–372 (2015).

148. Joensuu, H. *et al.* Risk of recurrence of gastrointestinal stromal tumour after surgery: an analysis of pooled population-based cohorts. *Lancet Oncol.* **13**, 265–274 (2012).

149. Han, C., Sun, X., Yang, Y., Che, Y. & Qin, Y. Brain Complex Network Characteristic Analysis of Fatigue during Simulated Driving Based on Electroencephalogram Signals. *Entropy* **21**, 353 (2019).

150. Ushenko, V. A. *et al.* Embossed topographic depolarisation maps of biological tissues with different morphological structures. *Sci. Rep.* **11**, 3871 (2021).

151. Bemis, K. D. *et al.* Cardinal: An R package for statistical analysis of mass spectrometry-based imaging experiments. *Bioinformatics* **31**, 2418–2420 (2015).

152. Verbeeck, N., Caprioli, R. M. & Van de Plas, R. Unsupervised machine learning for exploratory data analysis in imaging mass spectrometry. *Mass Spectrom. Rev.* **39**, 245–291 (2020).

153. Chung, H.-H., Huang, P., Chen, C.-L., Lee, C. & Hsu, C.-C. Next-generation pathology practices with mass spectrometry imaging. *Mass Spectrom. Rev.* e21795 (2022). doi:10.1002/mas.21795

154. Neumann, E., Romer, C., Allen, J. & Spraggins, J. Automatic Deposition of DAN Matrix using a TM Sprayer for MALDI Analysis of Lipids. *protocols.io* (2021).

155. Neumann, E., Allen, J., Anderson, D., Gutierrez, D. & Spraggins, J. High Resolution Imaging Mass Spectrometry Analysis using Bruker Daltonics Platforms. *protocols.io* (2019).

156. Patterson, N. H. *et al.* Next Generation Histology-Directed Imaging Mass Spectrometry Driven by Autofluorescence Microscopy. *Anal. Chem* **90**, 40 (2018).

157. Neumann, E. *et al.* PAS Staining of Fresh Frozen or Paraffin Embedded Human Kidney Tissue. *protocols.io* (2021).

158. Patterson, H. NHPatterson/napari-imsmicrolink: IMS MicroLink v0.1.7. (2022). doi:10.5281/ZENODO.6562052

159. Patterson, H. & Manz, T. NHPatterson/wsireg: wsireg v0.3.5. (2022). doi:10.5281/ZENODO.6561996

160. Fahy, E. *et al.* Update of the LIPID MAPS comprehensive classification system for lipids1. *J. Lipid Res.* **50**, S9–S14 (2009).

161. van der Walt, S. *et al.* scikit-image: image processing in Python. *PeerJ* **2**, e453 (2014).

162. Hunter, J. D. Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).