UNIVERSITY OF OKLAHOMA

GRADUATE COLLEGE

USING NOVEL DATA ANALYSIS METHODS TO EXTRACT INFORMATION FROM MASS SPECTROMETRY IMAGING AND SINGLE CELL MASS SPECTROMETRY RESULTS

A DISSERTATION

SUBMITTED TO THE GRADUATE FACULTY

in partial fulfillment of the requirements for the

Degree of

DOCTOR OF PHILOSOPHY

By

ZHU ZOU

Norman, Oklahoma

2023

USING NOVEL DATA ANALYSIS METHODS TO EXTRACT INFORMATION FROM MASS SPECTROMETRY IMAGING AND SINGLE CELL MASS SPECTROMETRY RESULTS

A DISSERTATION APPROVED FOR THE

DEPARTMENT OF CHEMISTRY AND BIOCHEMISTRY

By THE COMMITTEE CONSISTING OF

Dr. Zhibo Yang, Chair

Dr. Yihan Shao, Co-Chair

Dr. Daniel Glatzhofer

Dr. Ying Wang

This doctoral dissertation is dedicated to my family and friends.

Acknowledgments

First, I would like to express my deepest gratitude to my family for their limitless support throughout my graduate student career. Their encouragement and understanding have made this challenging time of my life much easier. I would not be who I am today without them.

Next, I would like to thank my committee members for guiding me to the right track. Dr. Yang, I still remember the day when you interviewed me for my admission before joining OU. Thank you for being my role model, expanding my horizon with your broad chemistry background and guiding me through the difficulties I came across in the past few years. I appreciate all the skills and knowledge you have shared with me, which have become the cornerstone of my projects. Dr. Shao, thank you for being my co-advisor and sharing the valuable experiences in your career, which inspired me a lot and helped me make up my mind. Thank you for all the encouragement when I felt lost and pointing out the possibilities to improve myself. I would like to thank Dr. Wang and Dr. Glatzhofer for being on my committee and sparing your precious time with me. I appreciate your valuable questions and feedback during the annual meetings guiding me towards the right track. I would also like to thank Dr. McCall for being both my committee member and collaborator before leaving OU.

Additionally, I would like to thank Dr. Foster, Dr. Yip, and other faculty members in the department who helped me a lot. I really appreciate your valuable suggestions on graduate student life, future career, and individual development.

Finally, I would like to thank my collaborators for bringing me the opportunities of all those wonderful projects and my lab members for helping me through my growth in my graduate student career, especially Dr. Xiang Tian and Dr. Genwei Zhang, who mentored me in the first few years, taught me the basics of mass spectrometry, and provided me with valuable advice so that I could advance in this field.

Table of Contents

List of figures

List of tables

Abstract

Mass spectrometry (MS) is a powerful tool for qualitative and quantitative biological sample analysis, with its high sensitivity, broad applicability, and strong robustness. While insights into ‘what’ and ‘how’ can be provided by mass spectrometry, MS technique coupled with traditional separation methods, such as liquid chromatography LC-MS, do not provide a satisfactory key to the question ‘where’ with high enough resolution. Accordingly, two different types of MS methods are being developed to analyze sample species with spatial resolution, one being single cell mass spectrometry (SCMS), and the other as mass spectrometry imaging (MSI).

In general, SCMS provides chemical insight of individual cells, whereas MSI can reveal the spatial distributions of chemical substances at a micrometer resolution. With the in-house microscale sampling device developed in the Yang group, the Single-probe, SCMS and MSI could be conducted. Both methods require specific protocols for sample treatment, experiment operation, data acquisition, and data analysis. Both SCMS and MSI studies are focused on analysis of small molecules (e.g., metabolites, lipids, and drug compounds). Different from SCMS studies, MSI measurements allow for acquiring spatial information on top of the chemical information provided by MS, and the spatial information has brought more complexity in the output data which require more advanced data analysis tools.

In this dissertation, the background of spatially resolved MS methods, i.e., MSI techniques, are first introduced, followed by a summary of previously published studies on quantitative SCMS metabolomics projects. In Chapter 3, MSI attempts on three different types of samples, including mice retina, patient breast, and co-cultured cancer cell spheroids, are introduced. In Chapter 4, the metabolomic profile changes of heart tissues upon *Trypanosoma Cruzi* infection have been investigated with the Single-probe MSI technique. The compatibility between MS and two commonly adopted strategies, fixation and staining, is studied, suggesting that X-gal staining has significantly altered the chemical profile. In Chapter 5, to handle SCMS data with better efficiency and higher mass accuracy, a Python-based MS data pretreatment platform with an easy-to-use graphical user interface (GUI) and an innovative peak alignment algorithm is developed to be compatible with improvised SCMS experiments. In Chapter 6, advanced statistical methods used for MS data processing are discussed in the context of MSI study of mice brain with Alzheimer's Disease as an example. The fusion of ion images from MSI and fluorescence images from immunohistology staining has improved the spatial resolution to a higher level, leading to more precise mapping of chemical substances and more findings involved in Alzheimer’s Disease development.

Chapter 1 Introduction

1.1. Background

In order to achieve better understanding of disease pathogenesis and biological activities, finer spatial resolution is required to break down the whole organism into multiple components at levels of organs, tissues, and cells.[1] Apart from MS-based methods, microscopy imaging[2], fluorescence imaging[3], magnetic resonance imaging[4], positron emission tomography[5], Raman spectroscopy imaging[6, 7], and photoacoustic imaging[8] can all provide biological or chemical information with spatial resolution. However, MS has the unique capability of detecting hundreds or even thousands of species at the same time without extra labeling, making it extremely powerful for bioanalytical tasks, such as untargeted analysis and biomarker discovery.[9, 10] Different strategies have been employed to obtain spatial information as MS data is being collected, with MSI and SCMS as the two major strategies and thus the main topics throughout this dissertation.

MSI was first conceptualized more than 50 years ago by Castaing and Slodzian, who proposed reading the sample plane as multiple periodically aligned pixels with mass spectrometers.[11] From MS scans from all pixels, a heatmap can be constructed for each extracted ion chromatogram, thus mapping the distribution of the selected

ions in the scanned region of interest in a visualizable and rather intuitive manner. The sample of MSI is typically a flat plane from either thin tissue sections or special cell cultures.[12, 13]

The MS analysis on single cells can be traced back to 1970s as well when Iliffe and coworkers studied the concentration of amino acid in individual neurons with MS.[14] Instead of scanning throughout the surface, most SCMS methods, on the other hand, rely on a successful isolation of single cells, usually with the assistance of sampling compartment with sharp focus or device with precise fluidic control, thus providing the spatial information of the chemical profile obtained from the mass spectra.[15, 16]

There have also been other MS-based methods, such as MasSpec Pen[17, 18] and chemical cartography,[19] that provide spatial information along with chemical profile, but the spatial resolution MasSpec Pen and chemical cartography can reach is lower than most MSI or SCMS methods. The size of MS data has also increased dramatically with higher throughput and instrument sensitivity. To efficiently analyze MS data, multiple software packages have been published with different algorithms and statistical methods.[20-22]

1.2. Categories of MS techniques

Today, MSI, SCMS and other MS studies have been reported by many different groups across the world. Depending upon the ionization method, two main categories can be defined: vacuum-based methods and ambient-based methods.

1.2.1. Vacuum-based method

1.2.1.1. Secondary Ion Mass Spectrometry (SIMS)

Secondary Ion Mass Spectrometry (SIMS) is an ionization method first developed in 1910, where ions were found to be generated from ion bombardment on a solid surface.[23, 24] Along with the development of vacuum systems and mass analyzers, SIMS became a more mature technique after 1970s when a series of studies on SIMS were conducted in the field.[25] By sputtering the sample surface with a primary ion beam (e.g. He^+, Ne^+, $C60^+$), secondary ions can be generated at the site of the bombardment and can be detected by the coupled mass analyzer and detector. The spatial resolution of SIMS can be as high as nanometer level with a proper focus of the primary ion beam, but the usage of SIMS on biological samples is so far hindered by the hard ionization causing fragmentation and thus difficulty in finding the nominal mass of the parent ion species.

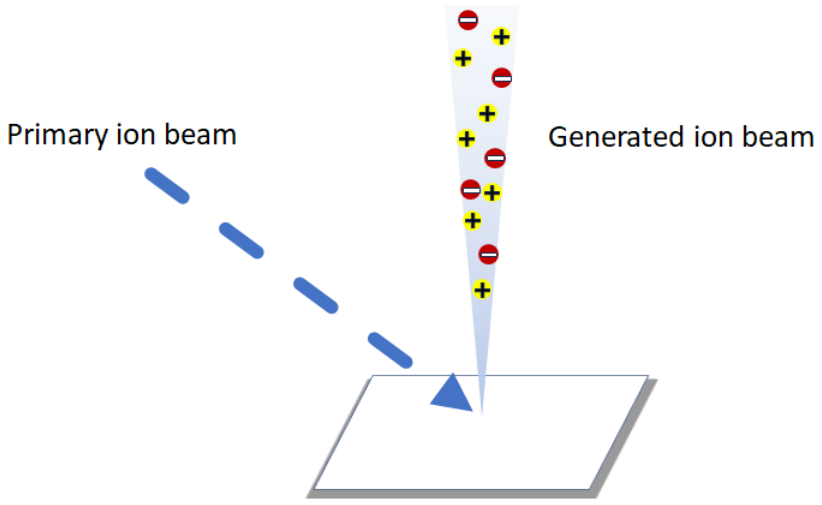

Figure 1-1 Schematic diagram of SIMS

1.2.1.2. Matrix-assisted Laser Desorption/Ionization (MALDI)

Matrix-assisted Laser Desorption/Ionization (MALDI) is nowadays one of the most popular vacuum-based ionization methods due to its high sensitivity and biocompatibility as a soft ionization technique.[26, 27] With MALDI, specific matrix molecules must be applied onto the sample surface prior to the experiment, which absorb energy from the laser beam and transfer the energy to other species for ionization. Depending on the specific target of the studies, pipelines with different matrices molecules have been established, such as 2,4,6-trihydroxyacetophenone (THAP) for lipids and α-Cyano-4-hydroxycinnamic acid (HCCA) for peptides.[26]

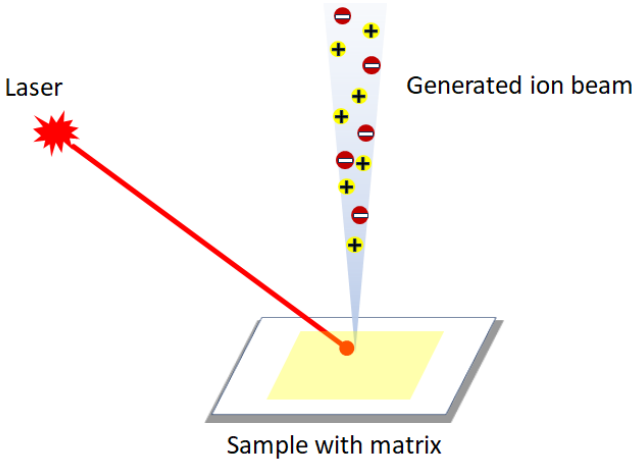

Figure 1-2 Schematic diagram of MALDI

1.2.2. Ambient method

1.2.2.1. Electrospray ionization (ESI)

Electrospray ionization (ESI) is another commonly used technique in mass spectrometry, which was first reported in 1968 by Malcolm Dole[28] It was later applied to biological macromolecules by John Bennett Fenn, who was awarded the Nobel Prize in Chemistry in 2002 for this contribution.[29] Despite the broad use of ESI, the theoretical basis of the ESI mechanism is yet to be fully determined. In the most accepted theory, the presence of high voltage and assistance of nebulizing gas prompted the sprayed, charged solvent droplets fragment into multiple smaller droplets due to overwhelming electrostatic repulsion against the surface tension. As

the fragmentation process repeats, the droplet degrades steadily until a mist of electrospray is formed.[30, 31]

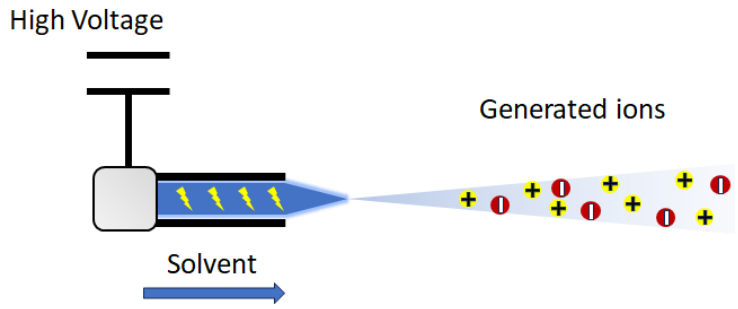

Figure 1-3 Schematic diagram of ESI

1.2.2.2. Desorption electrospray ionization (DESI)

Desorption electrospray ionization (DESI), which is the first type of MS technique for surface analysis in ambient environment, was developed by Graham Cooks' group.[32] In the DESI ionization process, an electrospray is generated to spray charged solvent droplets onto the surface of the sample, resulting in the transfer of energy and desorption of ionizable species from the sample.[33, 34]

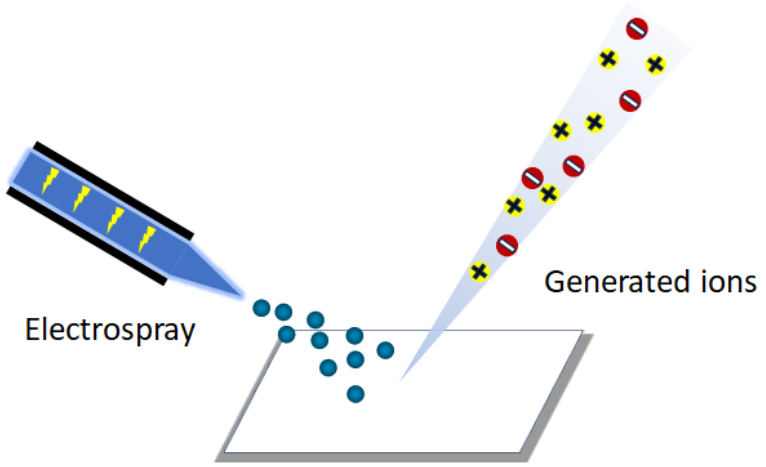

Figure 1-4 Schematic diagram of DESI

1.2.2.3. Laser ablation electrospray ionization (LAESI)

Laser ablation electrospray ionization (LAESI) is an ambient ionization technique developed in the Akos Vertes lab in 2007.[35] In the process of LAESI, analyte species blazed on the sample plane undergo desorption and get released into the air, then mixed, ionized, and carried to the MS inlet by the mist of electrospray flows over the sample surface.

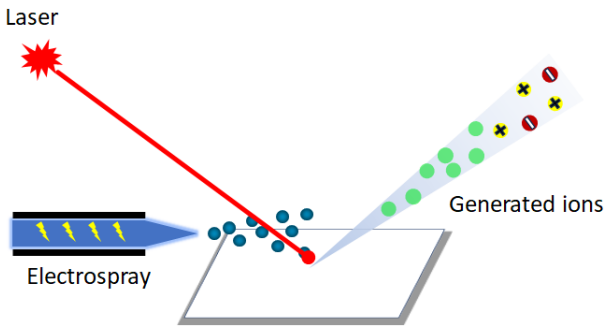

Figure 1-5 Schematic diagram of LAESI

1.2.2.4. Nano desorption electrospray ionization (Nano-DESI)

Nano desorption electrospray ionization (Nano-DESI) is a liquid-extraction based ionization method for surface sampling developed by the Julia Laskin group in 2010.[36, 37] Briefly, solvent flows through two connected capillaries and forms a liquid bridge at the junction between the two capillaries, where analytes from the surface are extracted and delivered to the MS inlet by the secondary capillary.

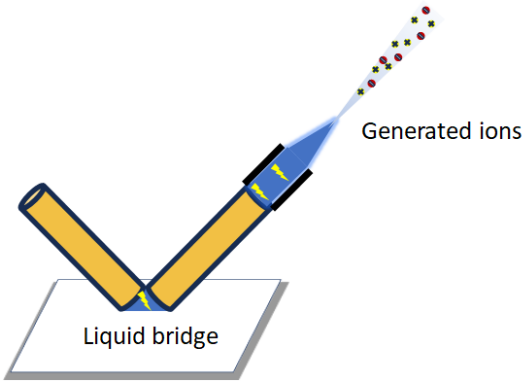

Figure 1-6 Schematic diagram of Nano-DESI

1.2.2.5. Single-probe

The Single-probe is a homemade micro-sampling device developed in the Zhibo Yang group in 2014.[38] The Single-probe consists of a laser-pulled dual-bore quartz tip and two capillaries (I.e., one is capillary providing solvent and the other one is a nano-ESI emitter) inserted into the two bores of the dual-bore quartz tip. Liquid junction for analyte extraction forms at the sharp tip of the probe, which can be as small as ~9-10 μm to sample a specific single cell or a pixel on the plane.[38, 39]

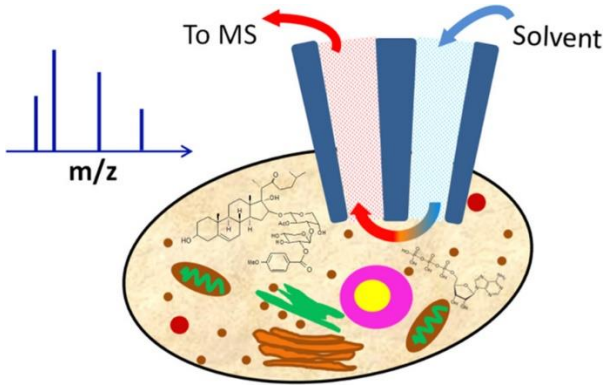

Figure 1-7 Schematic diagram of the Single-probe

Pan, N.; Rao, W.; Kothapalli, N. R.; Liu, R.; Burgett, A. W. G.; Yang, Z. *Analytical Chemistry* **2014**, *86* (19), 9376-9380.

1.3. Categories of sampling methods

According to the actual experimental setup, sampling of analyte could take place either before or during the ionization process. Spatial information is provided during ionization when the ionization process is guided with focused laser, ion beam, ion spray, or physical junction at designated spot.[39-43] While the ionization process does not provide any spatial information, special sampling strategies can be adopted to separate the analytes at different pixels or from different cells.

1.3.1. Sampling during ionization

SIMS, MALDI, DESI, LAESI, Nano-DESI and Single-probe are ionization methods that have capabilities to sample from a specific spot with guidance from focused ion beam, laser, electrospray, or physical tips of capillaries or probes.

1.3.2. Sampling before ionization

Although numerous ionization methods can provide spatial information during the occurrence of ionization, those methods may not work well for the actual system under study or match the original equipment of the laboratory. An additional sampling could be performed prior to the ionization process. For example, Single Cell Printer technology with Liquid Vortex Capture mass spectrometry (SCP-LVC-

MS) couples Single Cell Printer, a fluidic-based single cell separation device, followed by Liquid Vortex Capture and then electrospray ionization.[44] The Nonami group fabricated cell pressure probe with a sharp tip as 3-5 micrometers for sampling from a specific cell on a non-even sample surface, without cell separation, fixation, or sectioning as sample preparation.[45] The probe containing analyte could be relocated for injection towards mass spectrometer using electrospray ionization.

1.4. Applications

With the label-free and non-specific nature of mass spectrometry, different applications have been established in multiple fields depending upon the target study. For bioanalytical assays, the target for MS studies falls on two major categories: metabolites and proteins.

1.4.1. Application in metabolomics

In the field of mass spectrometry, metabolites usually refer to small molecules (<1500 Da in molecular weight) that are intermediate or end product of metabolism.[46] Compared with proteins, which have amino acid backbones, the analysis of metabolomics is usually complicated by a large number of potential candidates, which may render similar or even identical m/z values, in MS^1 and

complex fragmentation patterns in MS^2 measurements due to the variety in the nature of the sample, complexity of treatment, and experimental conditions.[47, 48] Large amoutns of SCMS metabolomics studies have been reported. For example, Yin et al. performed electroosmotic extraction with a custom-built sampling platform to study the abundance of flavonoids and glucose in *Allium cepa* cells.[49] Another micro-sampling device, the Single-probe, was applied on HCT-116 and HeLa cell lines to study the metabolomic change under irinotecan treatment.[50]

1.4.2. Application in proteomics

Protein is one of the basic building blocks for cells, tissues, and organs.[1] Despite the limited possibility of amino acid residues that composes the smallest unit of a protein, post translational modifications (PTMs) are commonly found in in vivo samples, which increase the complexity and diversity of protein structure by introducing different proteoforms.[51] Bogdan et al. developed Single Cell ProtEomics by Mass Spectrometry (SCoPE-MS) system in 2018 to study the relationship between mRNA and protein levels in single mammalian cells, followed by the upgraded version SCoPE2 in 2021.[52, 53]

1.5. Currently methods and published tools for MS data analysis

Different software packages have been built to handle MS data, including large applications designed for multiple purposes such as MZmine[21, 54, 55], or a specific function for a certain step in the whole workflow such as imzMLConverter.[56] Given numerous options (e.g., whether chromatographic separation is included in the experiment, whether the acquired spectra is based upon profile or centroid, whether the analysis is targeted or untargeted, whether the analytes are protein or small molecules, and other options that may be unique to actual experimental setup), it is impossible to summarize and present a universal data analysis workflow for all types of MS experiments. However, as the general purpose of mass spectrometry measurements are essentially the same, strategies including noise filter, peak fitting, background removal, signal normalization, peak alignment, structural identification are commonly used.[57, 58] A few examples of currently published/commercialized software for MS data analysis are listed here.

1.5.1. Liquid chromatography tandem mass spectrometry data analysis

Liquid chromatography–mass spectrometry (LC/MS) is a widely used method in analytical chemistry.[59] It combines the capability of chromatographic separation with the powerful detection of mass spectrometry. The separation of analytes converts complicated mixtures into individual components, thus significantly

reduces the difficulty in analysis. OpenMS is an open-source LC/MS data analysis framework based on C++ and Python.[60] Functions including file handling, signal processing, identification and quantification were made available in the TOPP (The OpenMS Proteomics Pipeline).[61] XCMS is another open-source software available online designed for LC/MS-based data analysis which brings up an advantage as nonlinear retention time correction for LC/MS peaks.[20, 62]

1.5.2. Mass spectrometry imaging data analysis

MSI, which is introduced in Section 1.1, is another emerging field under rapid growth with an urgent need for the development of analysis packages for data processing and visualization. MSI QuickView is a software that generates 2D ion images with RAW data without prior format conversion, which significantly increases the efficiency for MSI data processing.[63] Cardinal is another package built in R for MSI data analysis, with a unique unsupervised image segmentation method named spatial shrunken centroids.[64, 65]

1.5.3. Structural identification

After MS data is generated and treated, connections must be made between the processed peaks and actual chemical substances before answers to biological

questions can be obtained. To identify the structure of peaks detected, two different strategies are commonly adopted. Based upon the m/z values of the precursor and fragment ions, the potential structures of the original compound can be derived for identification. In the field of proteomics, this method for identifying peptide sequence is called *de novo* sequencing with numerous published tools including PEAKS, PepNovo, and NovoHMM.[66-68] The other strategy is searching through the queries for the known substance in the database that matches best with the observed features. Mascot and MaxQuant are two examples of popular protein database search software packages.[22, 69] Besides proteomics usages, there are also databases for metabolomics studies such as METLIN and Human Metabolome Database (HMDB).[46, 70] Hybrid methods such as PEAKS DB and machine learning assisted method have also been studied to further improve the efficiency of structural identification.[71, 72]

1.6. References

(1) Farley, A.; McLafferty, E.; Hendry, C. Cells, tissues, organs and systems. *Nurs Stand* **2012**, *26* (52), 40-45. DOI: 10.7748/ns2012.08.26.52.40.c9248.
(2) Deshpande, N. M.; Gite, S.; Aluvalu, R. A review of microscopic analysis of blood cells for disease detection with AI perspective. *PeerJ Comput Sci* **2021**, *7*, e460. DOI: 10.7717/peerj-cs.460.
(3) Ormachea, O.; Villazón, A.; Rodriguez, P.; Zimic, M. A Smartphone-Based Low-Cost Inverted Laser Fluorescence Microscope for Disease Diagnosis. *Biosensors (Basel)* **2022**, *12* (11). DOI: 10.3390/bios12110960.
(4) Battineni, G.; Hossain, M. A.; Chintalapudi, N.; Traini, E.; Dhulipalla, V. R.; Ramasamy, M.; Amenta, F. Improved Alzheimer's Disease Detection by MRI Using Multimodal Machine Learning Algorithms. *Diagnostics (Basel)* **2021**, *11* (11). DOI: 10.3390/diagnostics11112103.
(5) Phelps, M. E. Positron emission tomography provides molecular imaging of biological processes. *Proc Natl Acad Sci U S A* **2000**, *97* (16), 9226-9233. DOI: 10.1073/pnas.97.16.9226.
(6) Zhang, Y.; Ren, L.; Wang, Q.; Wen, Z.; Liu, C.; Ding, Y. Raman Spectroscopy: A Potential Diagnostic Tool for Oral Diseases. *Front Cell Infect Microbiol* **2022**, *12*, 775236. DOI: 10.3389/fcimb.2022.775236.
(7) Kann, B.; Offerhaus, H. L.; Windbergs, M.; Otto, C. Raman microscopy for cellular investigations—From single cell imaging to drug carrier uptake visualization. *Advanced drug delivery reviews* **2015**, *89*, 71-90.
(8) Bayer, C. L.; Luke, G. P.; Emelianov, S. Y. PHOTOACOUSTIC IMAGING FOR MEDICAL DIAGNOSTICS. *Acoust Today* **2012**, *8* (4), 15-23. DOI: 10.1121/1.4788648.
(9) Olshina, M. A.; Sharon, M. Mass Spectrometry: A Technique of Many Faces. *Q Rev Biophys* **2016**, *49*. DOI: 10.1017/s0033583516000160.
(10) Glish, G. L.; Vachet, R. W. The basics of mass spectrometry in the twenty-first century. *Nature Reviews Drug Discovery* **2003**, *2* (2), 140-150. DOI: 10.1038/nrd1011.
(11) Castaing, R.; Slodzian, G. Microanalyse par émission ionique secondaire. *J Microsc* **1962**, *1* (1), 395-410.
(12) Buchberger, A. R.; DeLaney, K.; Johnson, J.; Li, L. Mass spectrometry imaging: a review of emerging advancements and future insights. *Analytical chemistry* **2018**, *90* (1), 240.
(13) Bien, T.; Koerfer, K.; Schwenzfeier, J.; Dreisewerd, K.; Soltwisch, J. Mass spectrometry imaging to explore molecular heterogeneity in cell culture. *Proc Natl Acad Sci U S A* **2022**, *119* (29), e2114365119. DOI: 10.1073/pnas.2114365119.

(14) Iliffe, T. M.; McAdoo, D. J.; Beyer, C. B.; Haber, B. Amino acid concentrations in the Aplysia nervous system: neurons with high glycine concentrations. *J Neurochem* **1977**, *28* (5), 1037-1042. DOI: 10.1111/j.1471-4159.1977.tb10666.x.
(15) Liu, R.; Pan, N.; Zhu, Y.; Yang, Z. T-Probe: An Integrated Microscale Device for Online In Situ Single Cell Analysis and Metabolic Profiling Using Mass Spectrometry. *Anal Chem* **2018**, *90* (18), 11078-11085. DOI: 10.1021/acs.analchem.8b02927.
(16) Abouleila, Y.; Onidani, K.; Ali, A.; Shoji, H.; Kawai, T.; Lim, C. T.; Kumar, V.; Okaya, S.; Kato, K.; Hiyama, E.; et al. Live single cell mass spectrometry reveals cancer-specific metabolic profiles of circulating tumor cells. *Cancer Sci* **2019**, *110* (2), 697-706. DOI: 10.1111/cas.13915.
(17) Sans, M.; Zhang, J.; Lin, J. Q.; Feider, C. L.; Giese, N.; Breen, M. T.; Sebastian, K.; Liu, J.; Sood, A. K.; Eberlin, L. S. Performance of the MasSpec Pen for Rapid Diagnosis of Ovarian Cancer. *Clinical Chemistry* **2019**, *65* (5), 674-683. DOI: 10.1373/clinchem.2018.299289 (acccessed 11/17/2023).
(18) Zhang, J.; Rector, J.; Lin, J. Q.; Young, J. H.; Sans, M.; Katta, N.; Giese, N.; Yu, W.; Nagi, C.; Suliburk, J.; et al. Nondestructive tissue analysis for ex vivo and in vivo cancer diagnosis using a handheld mass spectrometry system. *Sci Transl Med* **2017**, *9* (406). DOI: 10.1126/scitranslmed.aan3968.
(19) McCall, L.-I.; Morton, J. T.; Bernatchez, J. A.; de Siqueira-Neto, J. L.; Knight, R.; Dorrestein, P. C.; McKerrow, J. H. Mass Spectrometry-Based Chemical Cartography of a Cardiac Parasitic Infection. *Analytical Chemistry* **2017**, *89* (19), 10414-10421. DOI: 10.1021/acs.analchem.7b02423.
(20) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* **2006**, *78* (3), 779-787. DOI: 10.1021/ac051437y.
(21) Katajamaa, M.; Miettinen, J.; Orešič, M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **2006**, *22* (5), 634-636. DOI: 10.1093/bioinformatics/btk039 (acccessed 10/17/2023).
(22) Cox, J.; Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology* **2008**, *26* (12), 1367-1372. DOI: 10.1038/nbt.1511.
(23) Honig, R. E. The Growth of Secondary Ion Mass Spectrometry (SIMS): A Personal View of Its Development. In *Secondary Ion Mass Spectrometry SIMS V*, Berlin, Heidelberg, 1986//, 1986; Benninghoven, A., Colton, R. J., Simons, D. S., Werner, H. W., Eds.; Springer Berlin Heidelberg: pp 2-15.

(24) Thomson, J. LXXXIII. Rays of positive electricity. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1910**, *20* (118), 752-767.
(25) Haskins, N. J. MASS SPECTROMETRY | Liquid Secondary Ion Mass Spectrometry. In *Encyclopedia of Analytical Science (Second Edition)*, Worsfold, P., Townshend, A., Poole, C. Eds.; Elsevier, 2005; pp 380-386.
(26) Dreisewerd, K. Recent methodological advances in MALDI mass spectrometry. *Analytical and Bioanalytical Chemistry* **2014**, *406* (9), 2261-2278. DOI: 10.1007/s00216-014-7646-6.
(27) Darie-Ion, L.; Whitham, D.; Jayathirtha, M.; Rai, Y.; Neagu, A.-N.; Darie, C. C.; Petre, B. A. Applications of MALDI-MS/MS-based proteomics in biomedical research. *Molecules* **2022**, *27* (19), 6196.
(28) Dole, M.; Mack, L. L.; Hines, R. L.; Mobley, R. C.; Ferguson, L. D.; Alice, M. B. Molecular Beams of Macroions. *The Journal of Chemical Physics* **2003**, *49* (5), 2240-2249. DOI: 10.1063/1.1670391 (acccessed 11/17/2023).
(29) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989**, *246* (4926), 64-71. DOI: 10.1126/science.2675315.
(30) Konermann, L.; Ahadi, E.; Rodriguez, A. D.; Vahidi, S. Unraveling the Mechanism of Electrospray Ionization. *Analytical Chemistry* **2013**, *85* (1), 2-9. DOI: 10.1021/ac302789c.
(31) Iribarne, J. V.; Thomson, B. A. On the evaporation of small ions from charged droplets. *The Journal of Chemical Physics* **2008**, *64* (6), 2287-2294. DOI: 10.1063/1.432536 (acccessed 11/17/2023).
(32) Takats, Z.; Wiseman, J. M.; Gologan, B.; Cooks, R. G. Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science* **2004**, *306* (5695), 471-473.
(33) Morato, N. M.; Cooks, R. G. Desorption Electrospray Ionization Mass Spectrometry: 20 Years. *Accounts of Chemical Research* **2023**, *56* (18), 2526-2536. DOI: 10.1021/acs.accounts.3c00382.
(34) Balluff, B.; Hanselmann, M.; Heeren, R. M. A. Chapter Eight - Mass Spectrometry Imaging for the Investigation of Intratumor Heterogeneity. In *Advances in Cancer Research*, Drake, R. R., McDonnell, L. A. Eds.; Vol. 134; Academic Press, 2017; pp 201-230.
(35) Nemes, P.; Vertes, A. Laser ablation electrospray ionization for atmospheric pressure, in vivo, and imaging mass spectrometry. *Analytical chemistry* **2007**, *79* (21), 8098-8106.
(36) Roach, P. J.; Laskin, J.; Laskin, A. Nanospray desorption electrospray ionization: an ambient method for liquid-extraction surface sampling in mass spectrometry. *The Analyst* **2010**, *135* (9), 2233-2236. DOI: 10.1039/c0an00312c.

(37) Roach, P. J.; Laskin, J.; Laskin, A. Molecular Characterization of Organic Aerosols Using Nanospray-Desorption/Electrospray Ionization-Mass Spectrometry. *Analytical Chemistry* **2010**, *82* (19), 7979-7986. DOI: 10.1021/ac101449p.
(38) Pan, N.; Rao, W.; Kothapalli, N. R.; Liu, R.; Burgett, A. W. G.; Yang, Z. The Single-Probe: A Miniaturized Multifunctional Device for Single Cell Mass Spectrometry Analysis. *Analytical Chemistry* **2014**, *86* (19), 9376-9380. DOI: 10.1021/ac5029038.
(39) Rao, W.; Pan, N.; Tian, X.; Yang, Z. High-Resolution Ambient MS Imaging of Negative Ions in Positive Ion Mode: Using Dicationic Reagents with the Single-Probe. *Journal of the American Society for Mass Spectrometry* **2016**, *27* (1), 124-134. DOI: 10.1007/s13361-015-1287-7.
(40) Li, L.; Garden, R. W.; Romanova, E. V.; Sweedler, J. V. In Situ Sequencing of Peptides from Biological Tissues and Single Cells Using MALDI− PSD/CID Analysis. *Analytical Chemistry* **1999**, *71* (24), 5451-5458.
(41) Sheng, L.; Cai, L.; Wang, J.; Li, Z.; Mo, Y.; Zhang, S.; Xu, J.-J.; Zhang, X.; Chen, H.-Y. Simultaneous imaging of newly synthesized proteins and lipids in single cell by TOF-SIMS. *International Journal of Mass Spectrometry* **2017**, *421*, 238-244. DOI: https://doi.org/10.1016/j.ijms.2017.07.008.
(42) Wiseman, J. M.; Ifa, D. R.; Venter, A.; Cooks, R. G. Ambient molecular imaging by desorption electrospray ionization mass spectrometry. *Nature Protocols* **2008**, *3* (3), 517-524.
(43) Laskin, J.; Heath, B. S.; Roach, P. J.; Cazares, L.; Semmes, O. J. Tissue imaging using nanospray desorption electrospray ionization mass spectrometry. *Analytical chemistry* **2012**, *84* (1), 141-148.
(44) Cahill, J. F.; Riba, J.; Kertesz, V. Rapid, untargeted chemical profiling of single cells in their native environment. *Analytical chemistry* **2019**, *91* (9), 6118-6126.
(45) Gholipour, Y.; Erra-Balsells, R.; Hiraoka, K.; Nonami, H. Living cell manipulation, manageable sampling, and shotgun picoliter electrospray mass spectrometry for profiling metabolites. *Analytical Biochemistry* **2013**, *433* (1), 70-78.
(46) Wishart, D. S.; Tzur, D.; Knox, C.; Eisner, R.; Guo, A. C.; Young, N.; Cheng, D.; Jewell, K.; Arndt, D.; Sawhney, S.; et al. HMDB: the Human Metabolome Database. *Nucleic Acids Res* **2007**, *35* (Database issue), D521-526. DOI: 10.1093/nar/gkl923.
(47) Gowda, G. A.; Djukovic, D. Overview of mass spectrometry-based metabolomics: opportunities and challenges. *Methods Mol Biol* **2014**, *1198*, 3-12. DOI: 10.1007/978-1-4939-1258-2_1.

(48) Zhang, X. W.; Li, Q. H.; Xu, Z. D.; Dou, J. J. Mass spectrometry-based metabolomics in health and medical science: a systematic review. *RSC Adv* **2020**, *10* (6), 3092-3104. DOI: 10.1039/c9ra08985c.
(49) Yin, R.; Prabhakaran, V.; Laskin, J. Quantitative extraction and mass spectrometry analysis at a single-cell level. *Analytical chemistry* **2018**, *90* (13), 7937-7945.
(50) Pan, N.; Standke, S. J.; Kothapalli, N. R.; Sun, M.; Bensen, R. C.; Burgett, A. W.; Yang, Z. Quantification of drug molecules in live single cells using the single-probe mass spectrometry technique. *Analytical chemistry* **2019**, *91* (14), 9018-9024.
(51) Bennett, H. M.; Stephenson, W.; Rose, C. M.; Darmanis, S. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nature Methods* **2023**, *20* (3), 363-374. DOI: 10.1038/s41592-023-01791-5.
(52) Budnik, B.; Levy, E.; Harmange, G.; Slavov, N. SCoPE-MS: mass spectrometry of single mammalian cells quantifies proteome heterogeneity during cell differentiation. *Genome biology* **2018**, *19* (1), 1-12.
(53) Specht, H.; Emmott, E.; Petelski, A. A.; Huffman, R. G.; Perlman, D. H.; Serra, M.; Kharchenko, P.; Koller, A.; Slavov, N. Single-cell proteomic and transcriptomic analysis of macrophage heterogeneity using SCoPE2. *Genome biology* **2021**, *22* (1), 1-27.
(54) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**, *11* (1), 395. DOI: 10.1186/1471-2105-11-395.
(55) Schmid, R.; Heuckeroth, S.; Korf, A.; Smirnov, A.; Myers, O.; Dyrlund, T. S.; Bushuiev, R.; Murray, K. J.; Hoffmann, N.; Lu, M.; et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nature Biotechnology* **2023**, *41* (4), 447-449. DOI: 10.1038/s41587-023-01690-2.
(56) Race, A. M.; Styles, I. B.; Bunch, J. Inclusive sharing of mass spectrometry imaging data requires a converter for all. *Journal of Proteomics* **2012**, *75* (16), 5111-5112. DOI: https://doi.org/10.1016/j.jprot.2012.05.035.
(57) Chen, C.; Hou, J.; Tanner, J. J.; Cheng, J. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *Int J Mol Sci* **2020**, *21* (8). DOI: 10.3390/ijms21082873.
(58) Pendarvis, K.; Kumar, R.; Burgess, S. C.; Nanduri, B. An automated proteomic data analysis workflow for mass spectrometry. *BMC Bioinformatics* **2009**, *10* (11), S17. DOI: 10.1186/1471-2105-10-S11-S17.
(59) Pitt, J. J. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *Clin Biochem Rev* **2009**, *30* (1), 19-34.
(60) Sturm, M.; Bertsch, A.; Gröpl, C.; Hildebrandt, A.; Hussong, R.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Zerck, A.; Reinert, K.; et al. OpenMS – An open-

source software framework for mass spectrometry. *BMC Bioinformatics* **2008**, *9* (1), 163. DOI: 10.1186/1471-2105-9-163.
(61) Kohlbacher, O.; Reinert, K.; Gröpl, C.; Lange, E.; Pfeifer, N.; Schulz-Trieglaff, O.; Sturm, M. TOPP—the OpenMS proteomics pipeline. *Bioinformatics* **2007**, *23* (2), e191-e197. DOI: 10.1093/bioinformatics/btl299 (acccessed 11/17/2023).
(62) Forsberg, E. M.; Huan, T.; Rinehart, D.; Benton, H. P.; Warth, B.; Hilmers, B.; Siuzdak, G. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. *Nature Protocols* **2018**, *13* (4), 633-651. DOI: 10.1038/nprot.2017.151.
(63) Thomas, M.; Heath, B. S.; Laskin, J.; Li, D.; Liu, E.; Hui, K.; Kuprat, A. P.; van Dam, K. K.; Carson, J. P. Visualization of high resolution spatial mass spectrometric data during acquisition. *Annu Int Conf IEEE Eng Med Biol Soc* **2012**, *2012*, 5545-5548. DOI: 10.1109/embc.2012.6347250.
(64) Bemis, K. D.; Harry, A.; Eberlin, L. S.; Ferreira, C.; van de Ven, S. M.; Mallick, P.; Stolowitz, M.; Vitek, O. Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments. *Bioinformatics* **2015**, *31* (14), 2418-2420. DOI: 10.1093/bioinformatics/btv146 (acccessed 10/18/2023).
(65) Bemis, K. D.; Harry, A.; Eberlin, L. S.; Ferreira, C. R.; van de Ven, S. M.; Mallick, P.; Stolowitz, M.; Vitek, O. Probabilistic Segmentation of Mass Spectrometry (MS) Images Helps Select Important Ions and Characterize Confidence in the Resulting Segments. *Mol Cell Proteomics* **2016**, *15* (5), 1761-1772. DOI: 10.1074/mcp.O115.053918.
(66) Frank, A.; Pevzner, P. PepNovo: De Novo Peptide Sequencing via Probabilistic Network Modeling. *Analytical Chemistry* **2005**, *77* (4), 964-973. DOI: 10.1021/ac048788h.
(67) Fischer, B.; Roth, V.; Roos, F.; Grossmann, J.; Baginsky, S.; Widmayer, P.; Gruissem, W.; Buhmann, J. M. NovoHMM: A Hidden Markov Model for de Novo Peptide Sequencing. *Analytical Chemistry* **2005**, *77* (22), 7265-7273. DOI: 10.1021/ac0508853.
(68) Ma, B.; Zhang, K.; Hendrie, C.; Liang, C.; Li, M.; Doherty-Kirby, A.; Lajoie, G. PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry. *Rapid Commun Mass Spectrom* **2003**, *17* (20), 2337-2342. DOI: 10.1002/rcm.1196.
(69) Perkins, D. N.; Pappin, D. J.; Creasy, D. M.; Cottrell, J. S. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* **1999**, *20* (18), 3551-3567. DOI: 10.1002/(sici)1522-2683(19991201)20:18<3551::Aid-elps3551>3.0.Co;2-2.
(70) Guijas, C.; Montenegro-Burke, J. R.; Domingo-Almenara, X.; Palermo, A.; Warth, B.; Hermann, G.; Koellensperger, G.; Huan, T.; Uritboonthai, W.; Aisporna, A. E.; et al. METLIN: A Technology Platform for Identifying Knowns

and Unknowns. *Anal Chem* **2018**, *90* (5), 3156-3164. DOI: 10.1021/acs.analchem.7b04424.
(71) Zhang, J.; Xin, L.; Shan, B.; Chen, W.; Xie, M.; Yuen, D.; Zhang, W.; Zhang, Z.; Lajoie, G. A.; Ma, B. PEAKS DB: de novo sequencing assisted database search for sensitive and accurate peptide identification. *Mol Cell Proteomics* **2012**, *11* (4), M111.010587. DOI: 10.1074/mcp.M111.010587.
(72) Tran, N. H.; Qiao, R.; Xin, L.; Chen, X.; Liu, C.; Zhang, X.; Shan, B.; Ghodsi, A.; Li, M. Deep learning enables de novo peptide sequencing from data-independent-acquisition mass spectrometry. *Nature Methods* **2019**, *16* (1), 63-66. DOI: 10.1038/s41592-018-0260-3.

Chapter 2 State-of-Art Quantitative Single Cell Mass Spectrometry on Small Molecules

Section 2.3 of this chapter is included in a review manuscript with Yunpeng Lan (co-1st author) and Dr. Zhibo Yang, which is currently under review.

2.1. Abstract

To better understand the current state-of-art of quantitative single cell mass spectrometry (SCMS) methods adopted on small molecules, we collected information in multiple papers and summarized the methods for each study to learn how each group is handling the quantitation task differently. Methods extracted from different papers are described first, followed by discussion on how the methods can affect the data analysis process of SCMS data.

2.2. Introduction

In the field of mass spectrometry, so-called "bulk analysis" has been the golden standard for biological sample analysis with LC-MS as a routine tool for its label-free nature, high sensitivity, wide applicability, capability of quantitation and high

reproducibility. However, traditional bulk analysis has its limitations and shortcomings in many different aspects. SCMS results generated from individual single cells can provide insights into cell heterogeneity, which is an important concept that leads to more detailed pictures and reasonable explanations of biological activities. However, limited by the analyte amount of SCMS, strategies adopted in traditional bulk analysis such as LC-MS or GC-MS can hardly be directly applied, especially when the system was coupled with a sophisticated sampling device for single cell isolation. Numerous efforts have also been made to obtain quantitative information for chemical substances instead of only qualitative analysis. Not only the detail numbers can provide more information and help us catch subtle chemical profile differences, but also give us more understanding on subpopulations within a certain group, which provides more biological information, especially in comparison studies for marker discovery, disease pathogenesis, and drug metabolism studies. Especially for certain applications and diseases, significant changes, or patterns in the abundance of chemical species were enough to draw conclusions for diagnosis or other purposes. Despite the rich information from quantitative analysis, there has been increased difficulty compared with qualitative analysis due to higher requirement for sensitivity, system stability, and introduction of standard. Hereby we introduce a collection of quantitative SCMS studies based upon their normalization strategies, followed by summary and discussions.

2.3. Quantitative SCMS examples

2.3.1. Relative quantification without internal standard

Zhang *et al.* have coupled two different methods, droplet extraction and pulsed direct current electrospray ionization (pico-ESI) together, for single-cell analysis.[1] *In situ* single-cell analysis usually has the cell samples submerged in culture medium, which introduces a severe matrix effect interfering with MS detection at the single-cell level. Droplet extraction is a probe-based sampling method which is capable of minimizing the matrix effect of MS detection.[2] Cells were first washed with cold washing solution to remove culture medium, and the biochemical reactions in cells were weakened. A vacuum drying oven was used to dry the washed cells and localize cell contents on the surface of the cell culture dish. An emitter tip was used to soak the dried cells with extraction solvent, suck the extracts, and deliver to MS with the assisted solvent. During the ionization process, pico-ESI was adopted by connecting a DC electrode at the rear of the emitter. With the pulsed voltage applied onto the electrode and careful tuning, pulsed electrospray was established as pico-ESI.[3] The extended ion signal duration (up to 100 times) allowed for acquiring more MS2 spectra for molecular identification.

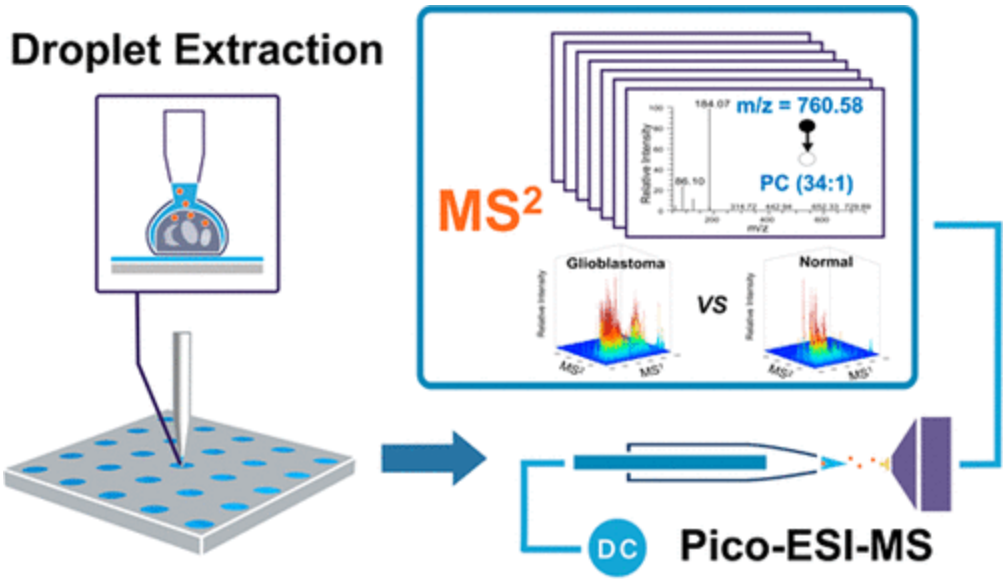

Figure 2-1 Schematic diagram of droplet extraction and pico-ESI.

Zhang, X.-C.; Zang, Q.; Zhao, H.; Ma, X.; Pan, X.; Feng, J.; Zhang, S.; Zhang, R.; Abliz, Z.; Zhang, X. *Analytical chemistry* **2018**, *90* (16), 9897-9903.

The stability of this method was tested using glucose and isotopically labelled glucose prior to studying human glioblastoma cell line (A172) and normal human astrocyte cell (HA) lines. Experimental results indicated that these two cell lines were differentiated on the PCA plot, with glioblastoma cells associated with a higher degree of PC desaturation. A potential biomarker for the cancer cell line, PE(20:1), was also identified in MS2 spectra. In addition, m/z 508.3398 was

identified as PC(17:1) in A172 cell line only, whereas this peak was contributed from two isobaric ions, PC(17:1) and PE(20:1), in HA cells.

Nakashima *et al.* conducted single-cell metabolites' profiling using the pressure probe electrospray ionization-mass spectrometry (IEC-PPESI-MS).[4] The cell pressure probe was originally developed to measure cellular properties, including turgor pressure and cell volume, of plant cells. The cell pressure probe has a sharp tip (3-5 μm), which is pre-filled with silicone oil, for precise sampling from a specific single cell. Using a microscope, images of tip geometry and position changes of meniscus, which is between the cell sap and silicone oil inside the tip, were captured to calculate sample volumes. Coupled to a mass spectrometer, a titanium wire electrode was inserted into the quartz capillary tip for ionization and detection of cellular analytes.

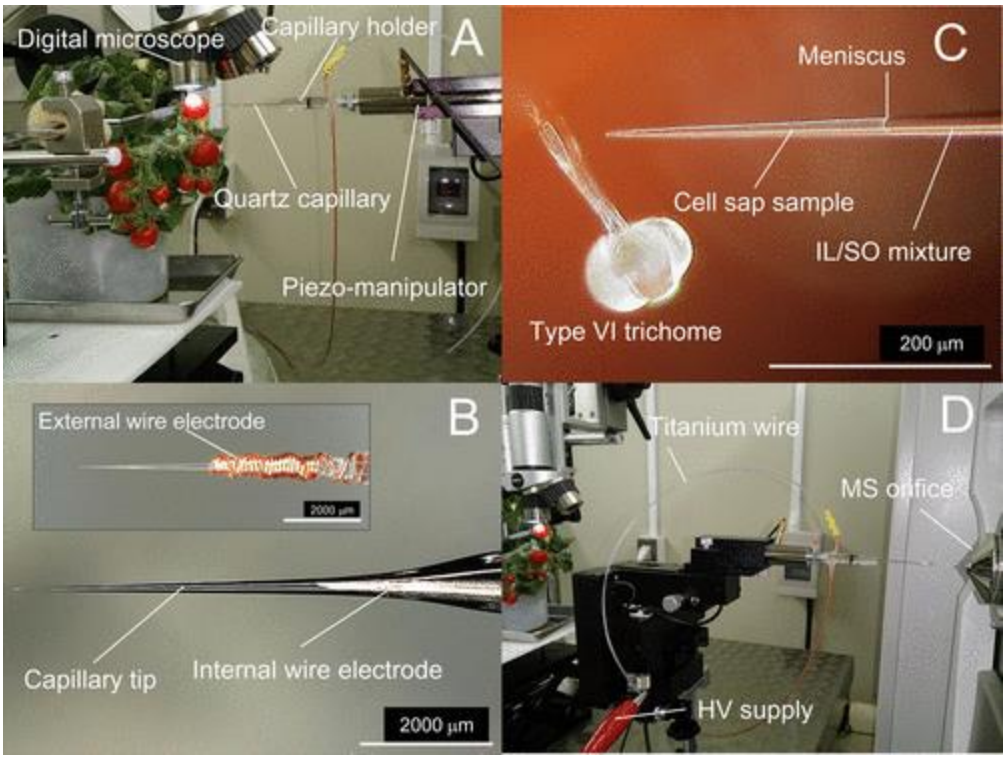

Figure 2-2 Experimental setup of IEC-PPESI-MS.

Nakashima, T.; Wada, H.; Morita, S.; Erra-Balsells, R.; Hiraoka, K.; Nonami, H. *Analytical Chemistry* **2016**, *88* (6), 3049-3057.

In this study, the pressure probe was used to sample a small amount of sap directly from stalk cells of tomato cultivar "Micro-tom". As the volume extracted from each cell is very small, adjacent cells can be analyzed in one batch. The lower limit of detection of several standard metabolites was reported as femtomole level, with a higher sensitivity in the negative ion mode due to lower background peaks. For

example, the lower limit of detection of ascorbic acid was dropped down to 1 and 0.01 femtomoles for positive and negative ion mode, respectively. Compared with the design with an external electrode, the one with an internal electrode increased the detection sensitivity of ascorbic acid by 32000 and 4000 times under the positive and negative ion modes, respectively. Differences in metabolite composition of cells from different trichomes were discovered, especially in flavonoids and acyl sugars. Molecular compositions were quantified based on relative peak intensity.

Abouleila *et al.* conducted live SCMS analyses of circulating tumor cells (CTC).[5] CTCs are considered to be directly related to cancer metastasis, and studies of this type of rare cells are crucial for cancer diagnosis and treatment. Blood samples from gastric cancer (GC) and colorectal cancer (CRC) patients were collected for single cell analysis. Red blood cells were first removed from the blood samples with red blood cells lysis buffer. The remaining cells, including CTCs and lymphocytes, were isolated by the ClearCell FX system prior to being transferred to an imaging petri dish. Cells are separated through Dean Flow Fractionation in curved channels. Separated single cells were then picked up by nanospray tips under a microscope and then kept in –80 °C environment until subsequential metabolite extraction and MS analysis. PCA-DA (principal component analysis followed by discriminant analysis) of MS data was carried out, and the results indicated clear separations between different cell types. 155 different species showed significant differences

between CTCs and lymphocytes (with log2-fold change > 1 and $p < 0.05$). Among these 155 species, acylcarnitine metabolites, sterol lipids, and eicosanoids were more abundant in CRC, whereas glycerophospholipids showed higher abundances in GC CTCs.

2.3.2. Relative quantification with internal standards

The Masujima group coupled 3D holographic and tomographic (HT) laser microscopy with live single-cell mass spectrometry to obtain mass spectra with 3D spatial resolution.[6] In their studies, human hepatocellular carcinoma cells (HepG2) were placed under the HT laser microscope. With a micromanipulator, a small volume of cytoplasm was extracted using a nanospray tip. The extraction volume was estimated according to the refractive index change in HT microscopy images captured from different angles. The spatial information of the extract was also monitored using the HT microscope. The nanospray tip was then coupled to an Orbitrap Velos Pro mass spectrometer for analysis. To quantify one of the detected species, methionine sulfoxide, a stable isotopically labeled internal standard, L-histidine-$^{15}N_3$, was added in the ionization solvent. The trapped volume in each measurement was estimated at around 1 pL ± 11 aL. Using a calibration curve, the concentration of methionine sulfoxide in the extract was estimated as 5.1 pmol/mL.

The Nonami group has also used the cell pressure probe for MS studies of tulip cells.[7] In this particular study, the pressure probe was coupled with an external electrode to apply high voltage to the analyte for electrospray ionization. A number of species (e.g., sugars, amino acids, vitamins, fatty acids, and secondary metabolites) were detected from different types of tulip cells. Particularly, metabolites related to the Krebs cycle were investigated in more detail. To obtain more accurate concentrations of the substances, mannitol was added in the cell sap as the internal standard, and the relative abundances of other species were compared by normalizing their peak intensities to potassiated mannitol.

Kertesz group combined single-cell printer technology with liquid vortex capture mass spectrometry (SCP-LVC-MS) for quantitative single-cell analysis.[8] The isolated droplets containing only one cell were selected by SCP and then captured by the LVC probe. At the upper surface, the dropped cells were mixed with a solvent that causes cell rupture due to osmotic pressure.[9] Cell contents were released and transported towards the mass spectrometer for ionization within several seconds.

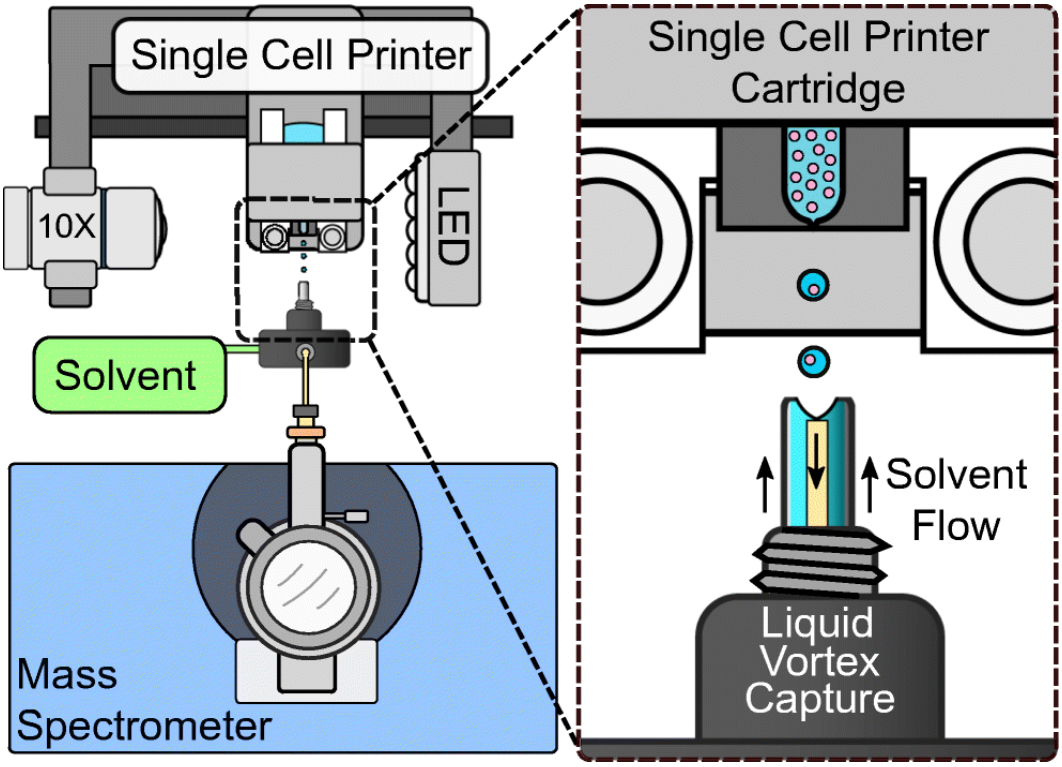

Figure 2-3 SCP-LVC-MS system with zoomed view at SCP-LVC interface.

Cahill, J. F.; Riba, J.; Kertesz, V. *Analytical chemistry* **2019**, *91* (9), 6118-6126.

LVC has the potential to be coupled with different liquid-based ionization methods, and ESI has been adopted in this study. With this sampling method, a high throughput (>20 cells/minute) analysis has been achieved. For normalization, 1 nM of diacylglyceryltrimethylhomo-Ser(32:0) (DGTS(32:0)), a lipid not detected in single cells but shares great structural similarity with DGTS class lipids in cells, was added into the LVC solvent as the internal standard. The system was tested

using three different types of cells: *Euglena gracilis* (EuGr), *Chlamydomonas reinhardtii* (ChRe), and Hela cells. Untargeted chemical profiling was performed, and the difference in chemical compositions was clearly observed among three different types of cells, allowing for a rapid classification of cell types based upon specific lipids. Quantification of lipids in ChRe cells under nutrient deprivation treatment was performed, and it was discovered that nutrient conditions affected the abundances of DGTS(34:4), DGTS(34:3), DGTS(38:4), MGDG(34:7), pheophytin a, and chlorophyll a.

2.3.3. Absolute quantification with internal standards

Yin *et al.* developed a method using precise electroosmotic extraction for quantitative single-cell analysis.[10] The electroosmotic extraction was performed with a custom-built platform combined with two electrodes and one pulled sharp nanopipette.

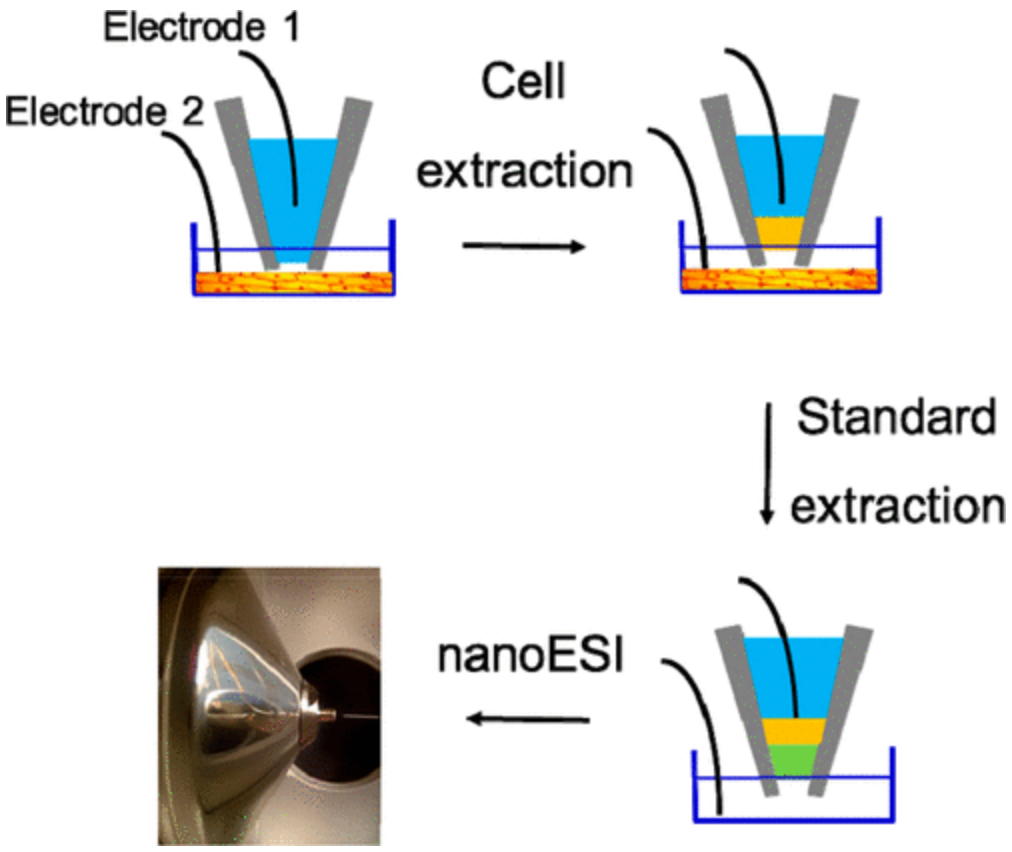

Figure 2-4 Schematic workflow of electroosmotic extraction.

Yin, R.; Prabhakaran, V.; Laskin, J. *Analytical chemistry* **2018**, *90* (13), 7937-7945.

Extracted volume was measured using images acquired by a high-resolution microscope. To perform quantitative analysis, 40 pL of 2 mM glucose-d2 solution (the internal standard) was also extracted by the device after the extraction of cytoplasm from *Allium cepa* cells. The nanopipette was then transferred to a mass spectrometer and used as a nanoESI emitter for MS analysis. More than 50

metabolites, including four flavonoids that were not previously reported, were detected from the 2-5 picolitres of extracted cytoplasm in the positive ion mode. Intensities of glucose and glucose-d2 were found to be well-correlated with their concentrations. It was discovered that, compared with other *in situ* SCMS analysis without separation, separation of compounds due to the mixing of water and hydrophobic electrolyte alleviated signal suppression.

Pedro and Rudewicz quantified the drug amiodarone (AMIO) and its metabolite, N-desethylamiodarone (NDEA), in single liver HepG2 and HepaRG cells.[11] In these studies, live cell 3D confocal microscopy imaging was coupled with high-resolution MS. To visualize different cellular components, cells were stained with three different fluorescent dyes: HCS LipidTOX Red phospholipidosis detection reagent (for phospholipids), Hoechst 33342 (for nucleus), and CellTracker Green CMFDA (for cytoplasm). A Yokogawa single-cell sampling system was used for whole-cell sampling. With the fluorescence label, cells ruptured during the sampling process could be easily ruled out. Each intact cell was isolated under the guidance of the confocal microscope, sucked into a platinum-coated glass tip (10 μm ID), completely lysed by methanol and water, dried by a heating block (70-95 °C), redissolved by methanol/water containing the internal standard (AMIO-D4 and NDEA-D4) with a known concentration, and then sprayed into the mass spectrometer. Calibration curves were established using a series of solutions of AMIO and NDEA containing their internal standards (AMIO-D4 and NDEA-D4)

at different concentrations. The volume of each cell was measured from sliced images using the 3D fluorescence confocal microscope. Concentrations of AMIO and NDEA in each cell were determined using the external calibration curve, amounts of AMIO and NDEA in each cell, and the volume of the analyzed cells. Using the interquartile range method, outliers were excluded from statistical analysis. 38 HepG2 cells and 31 HepaRG cells were analyzed to determine their intracellular AMIO and NDEA concentrations. NDEA concentration was found to be positively correlated with AMIO concentration in both types of cells, and a higher conversion ratio from AMIO to NDEA was found in HepaRG cells compared with HepG2 cells. Morphological changes were also studied in terms of the shape and volume of cells, phospholipid droplets, and nucleus.

The Yang group developed the Single-probe quantitative SCMS methods to determine the amounts and concentrations of intracellular drug compounds in single cells.[12, 13] The unique design of the Single-probe allows for flexible adoptions of sampling solutions with different compositions.

In these quantitative SCMS experiments, the internal standard (e.g., an isotopically labeled drug compound) with a known concentration was added in the sampling solution (e.g., acetonitrile or methanol/water). Both the target molecules (e.g., intracellular drug compounds) and their internal standard were simultaneously detected by MS. To measure the drug amount in adherent cells, a glass chip

containing microwells (diameter: 55 μm; depth: 25 μm) was used as a substrate in cell culture and drug treatment. The Single-probe was used to analyze single cells located inside individual wells, which minimized the diffusion loss of both intracellular drug and internal standard molecules during experiments. In these studies, HCT-116 and HeLa cell lines were used as models and treated by irinotecan with a series of treatment times and concentrations. The intracellular amounts of irinotecan were determined based on its peak areas relative to the internal standard (irinotecan-d10), the concentration of irinotecan-d10, flowrate of sampling solution, and data acquisition time. Comparative LC/MS studies were performed, and the results indicated that LC/MS experiments resulted in lower drug uptake, likely due to drug loss during sample preparation. To measure intracellular drug concentrations, the Single-probe SCMS setup was coupled with an integrated cell manipulation platform. T24 and K562 cells were treated by gemcitabine, rinsed by PBS, and resuspended in PBS. A single cell was captured by a cell-selection probe through a gentle suction from a connected microinjector, and the captured cell was transferred to the tip of the Single-probe by moving the manipulation platform. The single cell underwent a rapid lysis in the sampling solution containing the internal standard (13C,15N-labeled gemcitabine), followed by immediate MS analysis. The intracellular gemcitabine concentration in a single cell was determined from the measured drug amount, using the similar method as described above, and the cell

volume, which was measured using an inverted microscopy during single cell selection.

Circulating tumor cells (CTCs) have also been studied by Zhang *et al*.[14] Suspended CTCs were sampled with nanocapillaries and microcapillary holder with Ag/AgCl wire inserted for electro-osmotic extraction of cellular contents. In order to perform ease-to-use volumetric measurement during real sample analysis, experiments were conducted using a microscope to monitor the relationship between volume sucked into the tip and the electro-osmotic extration parameters (i.e., time and voltage). A linear relationship was found between extraction voltage and volume, and the combination of -2 V and 40 seconds was selected to acquire an extraction volume of ~120 fL. Two pairs of colorectal cancer cell lines were selected with different metastatic ability. Cell lysates were analyzed using LC-MS to screen meaningful metabolites. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis was carried out, and 14 metabolites from six different pathways were selected as the target analytes for the next-step single cell analysis. Isotopically labeled compounds were added as internal standards for quantification. With the observed cell heterogeneity and 11 quantified target metabolites, an unsupervised non-negative matrix factorization machine learning algorithm was adopted for clustering, and a 4-metabolite fingerprint classifier was built to divide CTCs into groups with different metastatic potentials.

2.3.4. Absolute quantification with standard addition

Rubakhin *et al.* used MALDI-MS for quantitative analysis of signaling peptides (SPs) in single neurons.[15] Neurons in the central nervous system contain highly varied abundances of SPs, which are critical for decoding brain functions. In these studies, tissues were harvested from *Aplysia californica*, and individual neurons were surgically isolated under a microscope. Single neurons were then dried on the sample plate and isotopically labeled with succinic anhydride or iTRAQ (isobaric tags for relative and absolute quantitation) for MALDI-MS studies for relative quantifications of SPs such as cerebrin and C_{β} peptides. In addition, absolute quantification of cerebrin was performed with standard addition. Briefly, the same spot is consecutively spiked with known amounts of cerebrin and re-analyzed after each addition. Each MALDI sampling process consumes a limited portion of the analyte, allowing for standard addition method for the same sample. The consistency of the signal from angiotensin I, which was added as an internal standard in the first step of sample preparation, validated the non-destructive nature of the sampling process. This is a very sensitive technique with a reported limit of detection and limit of quantification as 19 and 64 fmol, respectively. The experimental results indicated that an average of 230 fmol of cerebin were detected

in the F- and C-clusters compared with 700 fmol in ULAB nerves, which are in good agreement with the succinic anhydride labeling approach.

2.4. Results and discussion

Among the literature introduced above, four major categories based upon normalization strategies were listed including relative quantification without standard, relative quantification with internal standard, absolute quantification with internal standard, and absolute quantification with standard addition. When absolute quantification is being conducted, the standard compound must be an isotopically labelled version of the target analyte. Any differences in the chemical structure can possibly cause changes in ionization efficiency and thus lead to biased quantification result. For relative quantification, there is much more flexibility in the selection of standard compounds.

Different strategies have been applied for the identification of compounds in the samples. As the amount of analyte from individual single cells is highly limited, one is to mix a group of cells, while the other is to extend the signal duration of MS detection. However, given that MS/MS is not always available and reproducible, database matching is still the most popular method given by the limitation of single cell analysis.

2.5. Conclusion

Various methods for quantitative single cell mass spectrometry on small molecules are introduced and discussed in this chapter. From the summary, the workflow of single cell separation, mass spectrometry analysis, followed by data analysis can be a stereotype of the whole quantitative single cell mass spectrometry workflow. Given that MS/MS analysis is usually not available due to the limitation of experimental setup and low sample amount from individual cells, structural identification based upon MS1 database matching, which is the current dominating method, makes accurate mass detection and alignment a key factor towards efficient data mining and analysis, which is going to be the main topic for chapter 5.

2.6. References

(1) Zhang, X.-C.; Zang, Q.; Zhao, H.; Ma, X.; Pan, X.; Feng, J.; Zhang, S.; Zhang, R.; Abliz, Z.; Zhang, X. Combination of droplet extraction and Pico-ESI-MS allows the identification of metabolites from single cancer cells. *Analytical chemistry* **2018**, *90* (16), 9897-9903.
(2) Zhang, X.-C.; Wei, Z.-W.; Gong, X.-Y.; Si, X.-Y.; Zhao, Y.-Y.; Yang, C.-D.; Zhang, S.-C.; Zhang, X.-R. Integrated droplet-based microextraction with ESI-MS for removal of matrix interference in single-cell analysis. *Scientific reports* **2016**, *6* (1), 1-9.
(3) Wei, Z.; Xiong, X.; Guo, C.; Si, X.; Zhao, Y.; He, M.; Yang, C.; Xu, W.; Tang, F.; Fang, X. Pulsed direct current electrospray: enabling systematic analysis of small volume sample by boosting sample economy. *Analytical chemistry* **2015**, *87* (22), 11242-11248.
(4) Nakashima, T.; Wada, H.; Morita, S.; Erra-Balsells, R.; Hiraoka, K.; Nonami, H. Single-cell metabolite profiling of stalk and glandular cells of intact trichomes with internal electrode capillary pressure probe electrospray ionization mass spectrometry. *Analytical Chemistry* **2016**, *88* (6), 3049-3057.
(5) Abouleila, Y.; Onidani, K.; Ali, A.; Shoji, H.; Kawai, T.; Lim, C. T.; Kumar, V.; Okaya, S.; Kato, K.; Hiyama, E.; et al. Live single cell mass spectrometry reveals cancer-specific metabolic profiles of circulating tumor cells. *Cancer Sci* **2019**, *110* (2), 697-706. DOI: 10.1111/cas.13915.
(6) Ali, A.; Abouleila, Y.; Amer, S.; Furushima, R.; Emara, S.; Equis, S.; Cotte, Y.; Masujima, T. Quantitative live single-cell mass spectrometry with spatial evaluation by three-dimensional holographic and tomographic laser microscopy. *Analytical Sciences* **2016**, *32* (2), 125-127.
(7) Gholipour, Y.; Erra-Balsells, R.; Hiraoka, K.; Nonami, H. Living cell manipulation, manageable sampling, and shotgun picoliter electrospray mass spectrometry for profiling metabolites. *Analytical Biochemistry* **2013**, *433* (1), 70-78.
(8) Cahill, J. F.; Kertesz, V. Quantitation of amiodarone and N-desethylamiodarone in single HepG2 cells by single-cell printing-liquid vortex capture-mass spectrometry. *Anal Bioanal Chem* **2021**, *413* (28), 6917-6927. DOI: 10.1007/s00216-021-03652-6.
(9) Cahill, J. F.; Riba, J.; Kertesz, V. Rapid, untargeted chemical profiling of single cells in their native environment. *Analytical chemistry* **2019**, *91* (9), 6118-6126.
(10) Yin, R.; Prabhakaran, V.; Laskin, J. Quantitative extraction and mass spectrometry analysis at a single-cell level. *Analytical chemistry* **2018**, *90* (13), 7937-7945.

(11) Pedro, L.; Rudewicz, P. J. Analysis of live single cells by confocal microscopy and high-resolution mass spectrometry to study drug uptake, metabolism, and drug-induced phospholipidosis. *Analytical Chemistry* **2020**, *92* (24), 16005-16015.
(12) Pan, N.; Rao, W.; Kothapalli, N. R.; Liu, R.; Burgett, A. W. G.; Yang, Z. The Single-Probe: A Miniaturized Multifunctional Device for Single Cell Mass Spectrometry Analysis. *Analytical Chemistry* **2014**, *86* (19), 9376-9380. DOI: 10.1021/ac5029038.
(13) Pan, N.; Standke, S. J.; Kothapalli, N. R.; Sun, M.; Bensen, R. C.; Burgett, A. W.; Yang, Z. Quantification of drug molecules in live single cells using the single-probe mass spectrometry technique. *Analytical chemistry* **2019**, *91* (14), 9018-9024.
(14) Zhang, W.; Xu, F.; Yao, J.; Mao, C.; Zhu, M.; Qian, M.; Hu, J.; Zhong, H.; Zhou, J.; Shi, X.; et al. Single-cell metabolic fingerprints discover a cluster of circulating tumor cells with distinct metastatic potential. *Nature Communications* **2023**, *14* (1), 2485. DOI: 10.1038/s41467-023-38009-3.
(15) Rubakhin, S. S.; Sweedler, J. V. Quantitative measurements of cell− cell signaling peptides with single-cell MALDI MS. *Analytical chemistry* **2008**, *80* (18), 7128-7136.

Chapter 3 Special Sample Preparation for Mass Spectrometry Imaging Study: Retina, Breast, and Spheroids

This chapter involves three unpublished collaborative projects. The retina project was a collaboration with Dr. Pengchun Yu's group in Oklahoma Medical Research Foundation. After the sacrifice of mice and harvest of retina by Fei Han in the Yu group, the remaining work (cryo-sectioning, microscopy imaging, mass spectrometry) was conducted by me. The breast project was a collaboration with Dr. Zoran Gatalica in University of Oklahoma Health Sciences Center, where the breast biopsy was conducted. After that, cryo-sectioning, microscopy imaging, and mass spectrometry experiments were carried out by me. The spheroid project was a collaboration with Zongkai Peng in Dr. Zhibo Yang's group. Cell culture, spheroid culture, drug resistance induction and fluorescence imaging were jointly performed by me and Zongkai Peng. Cryo-section was carried out by me.

3.1. Abstract

As mass spectrometry imaging has become an emerging field of study, different samples were brought into the sight of researchers, including untypical samples that could be hardly treated with traditional MSI sample preparation methods due to

their physical properties. In this chapter, the MSI preparation for three main types of samples will be discussed, including in vivo samples such as mice retina, patient breast, and in vitro samples such as cancer cell spheroids.

3.2. Introduction

After the concept of mass spectrometry imaging has been introduced, different applications have been established on brain, liver, kidney tissues. A standard workflow has been established with embedding and snap freeze, which is widely applicable to many other types of tissues including heart and lung, sharing similar physical property. However, as the spatial resolution of mass spectrometry imaging has been pushed towards higher limits, different objects with all kinds of purposes have been brought up. Three specific samples will be discussed in this chapter.

Diabetic retinopathy is a common eye complication in diabetes patients.[1] Angiogenesis was found in the development of diabetic retinopathy, which leads to loss of vision and eventually blindness.[2] Despite the symptoms themselves being unlethal, the harmfulness is obvious and lifelong without available treatment to reverse or cure. Given the estimation that 9.60 million diabetic retinopathy cases were found among 36.32 million diabetes patients in the United States in 2021, there is an urge to better understand the development of this disease and to provide precautious treatment.[3] To better study the growth of blood vessels in retina, mice

were sacrificed when the blood vessel network in retina was still under development to capture the profile on the retinal surface.[4]

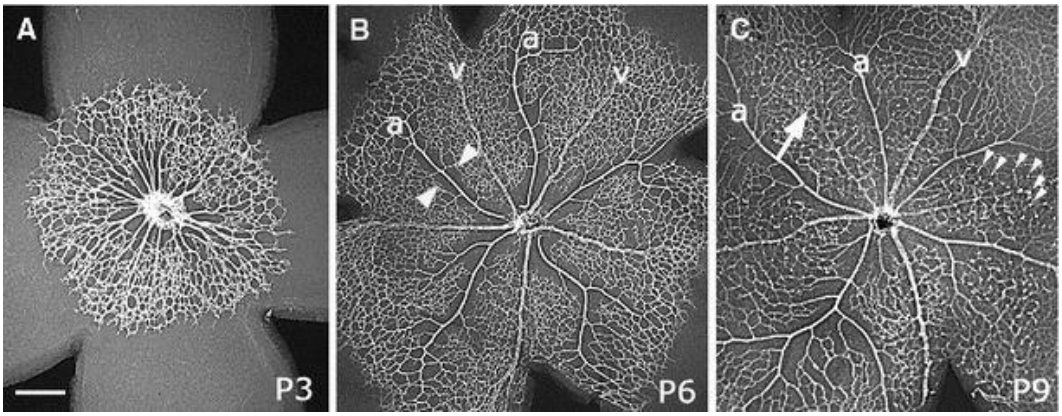

Figure 3-1 Development of mice retinal vasculature

Fruttiger, M. *Angiogenesis* **2007**, *10* (2), 77-88.

Breast cancer is the second most common cancer among women in the United States with nearly a quarter million new cases in 2020.[5] Metastasis can take place through blood vessels and lymphatic vessels in breast cancer patients, which increases the severeness of breast cancer.[6] There have been multiple different types of breast cancer. For example, triple-negative breast cancer (TNBC) is a type of breast cancer where the cancer cells do not have estrogen receptors, progesterone receptors, or overexpressed HER2 protein, and invasive ductal carcinoma is the

type of breast cancer which begins in the duct region and then spreads into other parts of the breast tissue.[7] Despite the diverse characteristics of different kinds of breast cancer, their specific chemical profile has not been thoroughly investigated yet.

Spheroid is a three-dimensional cell culture model which mimics the in vivo status of cells in patient bodies under in vitro culture conditions.[8] Compared with traditional cell cultures which spread over a plane and become homogenous across the surface, the spheroid has 3D scaffolds that are highly similar with in vivo tumors, providing cell heterogeneity induced due to the radial variation, thus can serve as a better in vitro model for cancer studies including tumor development and drug exposure.[9]

3.3. Methods

3.3.1. Retina sample treatment

Retina was carefully taken from eyeball harvested from mice right after sacrifice under a microscope and then immersed in 1x PBS to prevent possible morphology change due to dehydration. Cut was made onto the curved retinal surface to expand onto a flat surface formed by hydroxypropyl methylcellulose (HPMC) prior to cryo-sectioning.[10] Slides were taken for fluorescence imaging to locate the exact location of the transparent blood vessel cells under growth without blood cells. This was a

collaborative project with Pengchun Yu group in Oklahoma Medical Research Foundation and all mice protocols were approved by the relevant Institutional Animal Care and Use Committee.

3.3.2. Breast sample treatment

Breast tissue was obtained with surgical biopsy from patients followed by immediate storage in formalin solution for 24 hours for fixation purpose. The fixed tissue was taken from the fixative and rinsed with PBS, then embedded in HPMC for cryo-sectioning. All protocols were approved by OUHSC Institutional Review Board.

3.3.3. Spheroid sample treatment

Cancer cell spheroids were cultured using an established protocol in Yang group.[10] An adhesive human colorectal cancer cell line, HCT-116 was used. Cells were cultured in McCoy's 5A culture media with 10% FBS and 1% Pen Strep, incubated in 5% CO_2 and 37 °C in an incubator for growth prior to spheroid seeding. For co-culture spheroid samples, irinotecan (IR)-resistant HCT-116 cell line was developed by culturing wild type HCT 116 cells in irinotecan-rich culture medium. For differentiation, IR-resistant cell line was labeled with GFP (Green Fluorescence

Protein). Before seeding, the container for cancer cell spheroid formation must be specially treated to prevent cells from attaching to the container wall instead of each other. U-shape 96 well plates were selected and coated with gel formed from 1.7% agarose dissolved in basal medium. The first part of the seeding process was same as regular passaging of cells, including rinsing with PBS, trypsinization at 0.25% concentration, and trypsin deactivation with complete culture media. Next, a sip of the homogenized cell suspension was sampled and placed onto a hemacytometer to calculate the density of cells in the suspension. Each well of the 96-well plate was seeded with a total of about 10,000 – 12,000 cells with 200 µL complete cell growth medium, with 1:1 ratio between wild type cell line and IR-resistant cell line when seeding co-cultured spheroids. Medium was replaced every two days for spheroid formation and growth prior to harvesting. Harvested spheroids were first rinsed in PBS and then fixed with different fixatives of different choices, including formalin and glutaraldehyde, followed by embedding in 10% HPMC for cryo-sectioning at -20 °C.

3.4. Results and discussion

3.4.1. Retina sample results

Although mass spectrometry imaging has been conducted on retina samples in a perpendicular direction before, scanning in a parallel direction has not been

performed. As the blood vessels develop on the topmost layer of the retinal surface only, the result is extremely sensitive to the orientation of the sectioning process. As validation with other methods were not available due to the thickness of the region of interest in retinal surface, this project was discontinued.

3.4.2. Breast sample results

The major component of the breast tissue sample obtained was adipose, which was extremely soft. Unlike other common tissue samples for MSI studies that could be sectioned at -20 °C, breast tissue sample was not solidified as expected, thus cryo-section could not be performed as shapeshift was brought to the tissue when the touched by the blade. To increase the mechanical strength of the breast tissue, two methods were adopted including decreasing the cryo-section temperature and increasing the thickness of sectioning. Despite the increased mechanical strength of the breast tissue after the parameters were tuned, neither method could grant satisfactory sectioning condition individually. After the cryo-sectioning temperature reached -35 °C and the thickness of slides was increased to 40 µm, breast section was able to be made by the microtome blade. However, when the breast section was attached to a room temperature microscope slide, part of the tissue melted immediately under room temperature as the figure shows below.

Figure 3-2 Microscopy image of breast tissue section mounted on a microscope slide

Although part of the tissue remained intact as required, most of the tissue underwent shapeshift and delocalized, creating an uneven surface which was not compatible with MSI. The fusion of fat significantly damaged the original morphology of sectioned tissue. To explore the possibility of handling the situation, two methods

were adopted. First, as the breast tissue was not melted on the cryo-stat blade, attempts were made to keep both the tissue and the microscope slide at low temperature to prevent thawing. However, the tissue was unable to attach firmly due to the lack of temperature difference. As the rise in temperature could not be avoided due to the ambient environment of imaging experiments, another method was adopted by using absorbent TLC plate instead of regular glass slides.

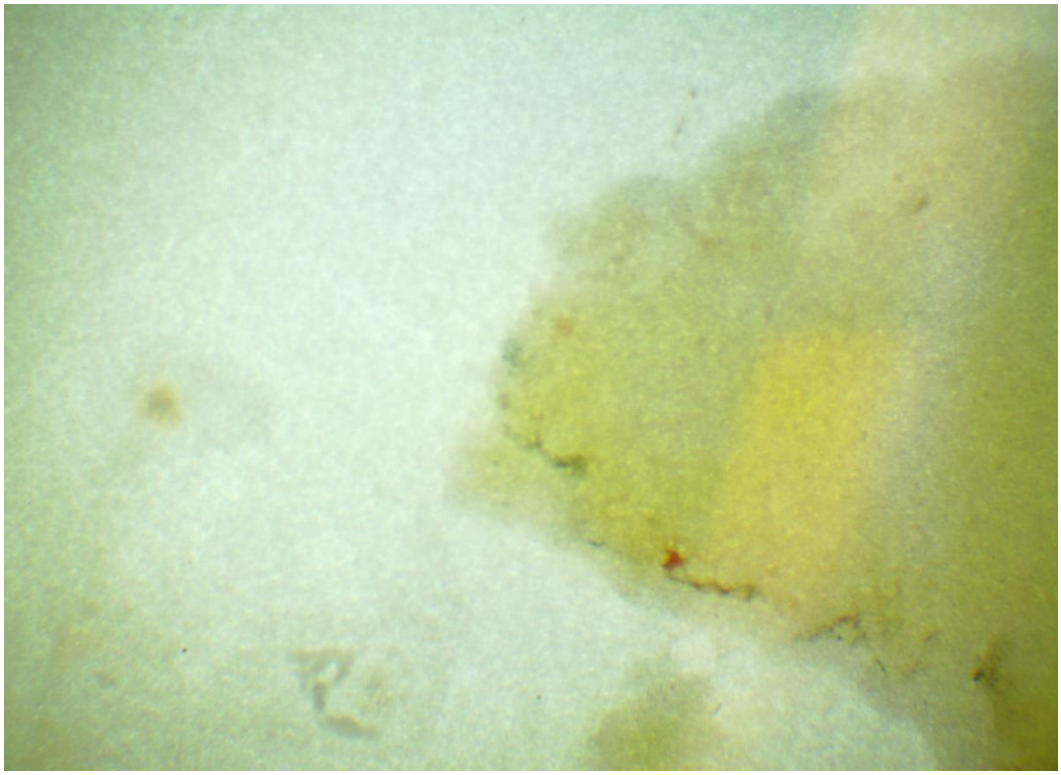

Figure 3-3 Microscope image of breast tissue section on a TLC plate

As shown in the microscope image above, the TLC plate solved the melting problem caused by the adipose tissue in the breast sample. However, absorbent material of the TLC plate brought new concern for MSI analysis. First, the smoothness of the TLC plate surface was not guaranteed, whereas height is one of the most sensitive parameters in MSI experiments. Change in the sampling height could result in an intensity drop or even absence of signal when the change was too obvious. Second, the solid particles coated on the TLC plate could be released into the air and be absorbed by the vacuum of mass spectrometers, which might potentially deal severe damage to the MS instrument. Third, the strong binding affinity of breast tissue towards the absorbent plate is much stronger than a regular glass slide, thus the performance of liquid-extraction based methods should be limited. With all the concerns above, the project was not continued.

3.4.3. Spheroid sample results

Co-culture spheroid was successfully grown and harvested with a similar protocol as mono-culture spheroid with the modification of seeding. Collected spheroids were successfully sectioned with desired thickness under regular conditions at -20 °C. To distinguish two types of HCT cells forming the co-culture spheroid, GFP-labeled HCT-116 cell line was adopted with differentiation between two cell lines using fluorescence. However, the fluorescence quickly quenched after collected

spheroids underwent cryo-section, which took them over 60 minutes out of their native growth environment on average before the fluorescence microscopy measurement. Fluorescence signal was found to be significantly decreased, thus unable to be used as a criterion for the differentiation between two cell lines. To keep the activity of the GFP expressed by one of the cell lines, fixation attempts were made using formaldehyde and glutaraldehyde. The fixation successfully enhanced the fluorescence intensity of the spheroid sections, with an estimated increase in signal of over 300%. Although fluorescence from GFP was reserved after fixation, irreversible morphology change was observed at the same time.

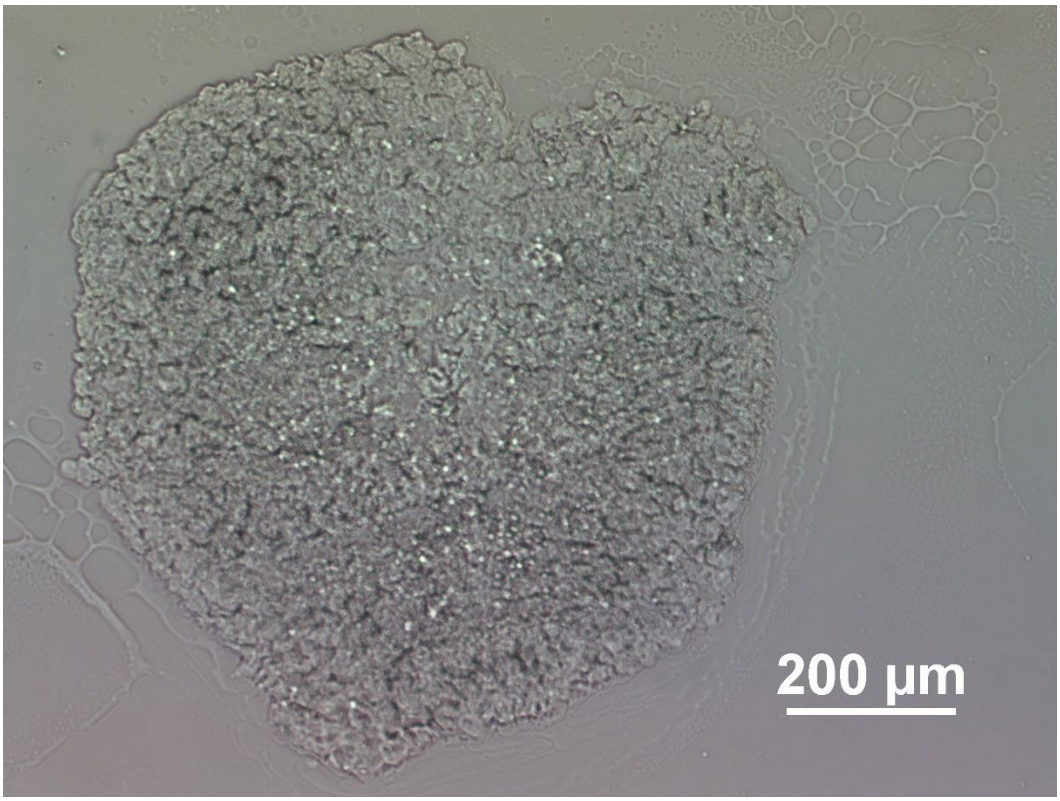

Figure 3-4 Bright field microscopy image of an unfixed spheroid section

This was an unfixed co-culture spheroid seeded with GFP labeled, IR resistant HCT-116 cell line and wild type HCT-116 cell line.

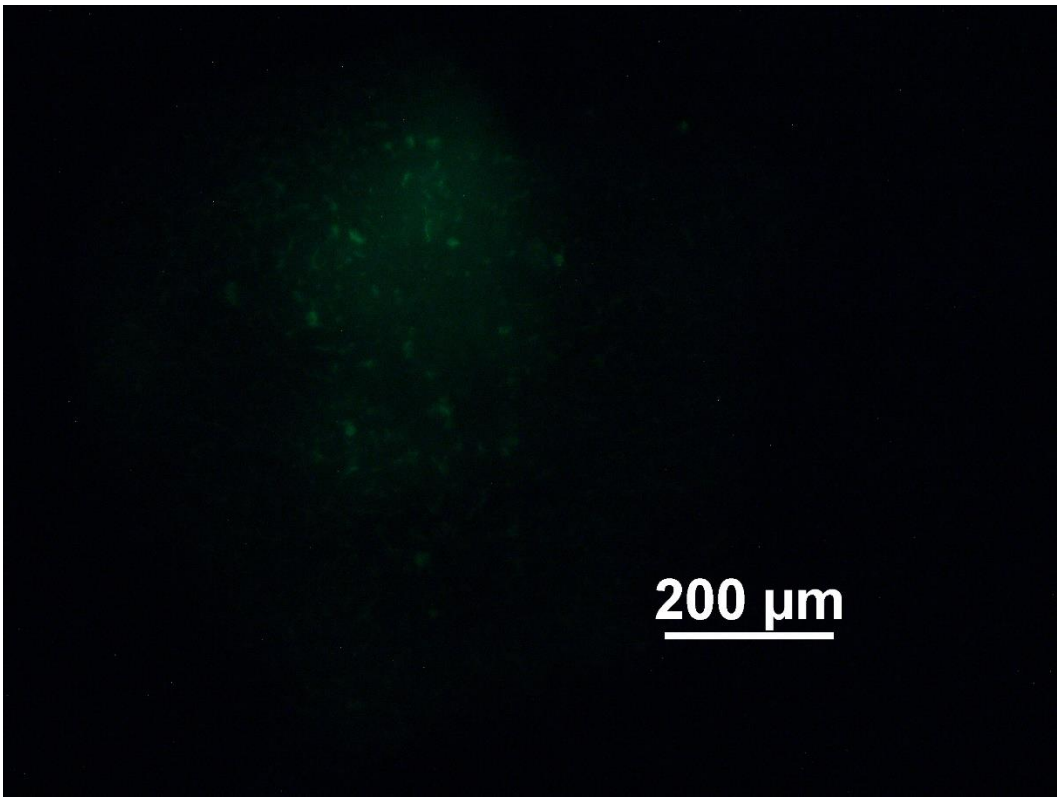

Figure 3-5 Fluorescence microscopy image of the same section

Section was from the same unfixed co-culture spheroid seeded with GFP labeled, IR resistant HCT-116 cell line and wild type HCT-116 cell line. The image was taken with 2 seconds as exposure and 6.80x as gain.

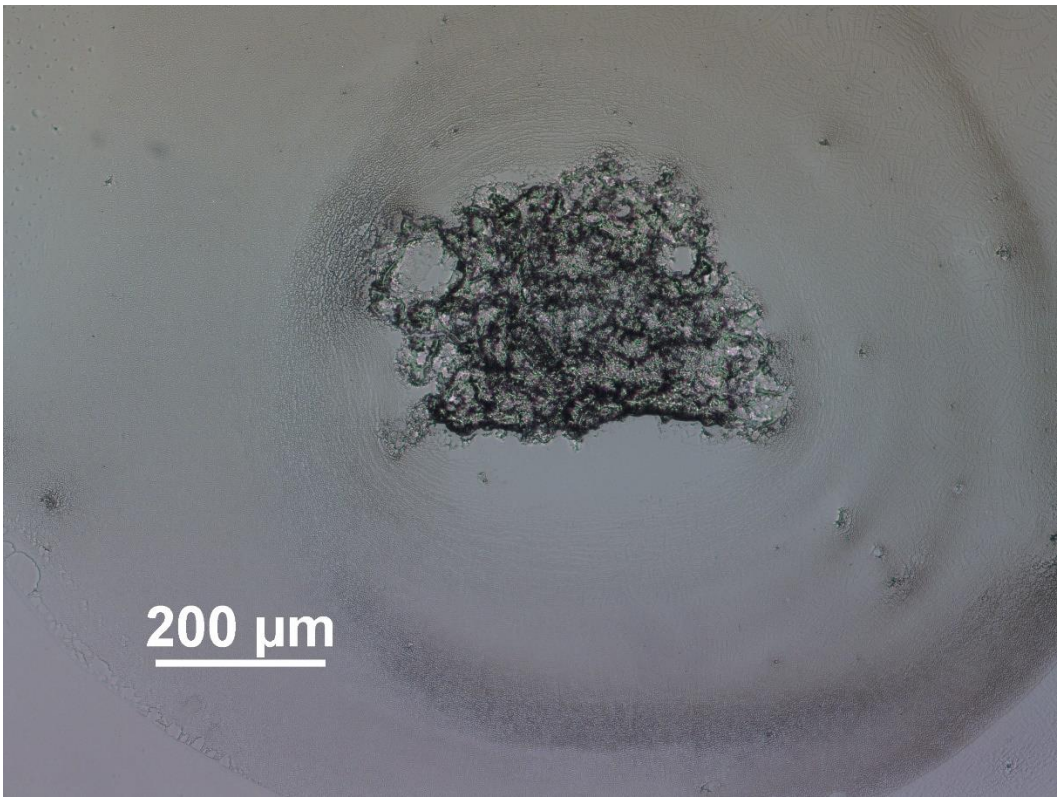

Figure 3-5 Bright field microscopy image of a fixed co-culture spheroid section

The spheroid was seeded with GFP labeled, IR resistant HCT-116 cell line and wild type HCT-116 cell line.

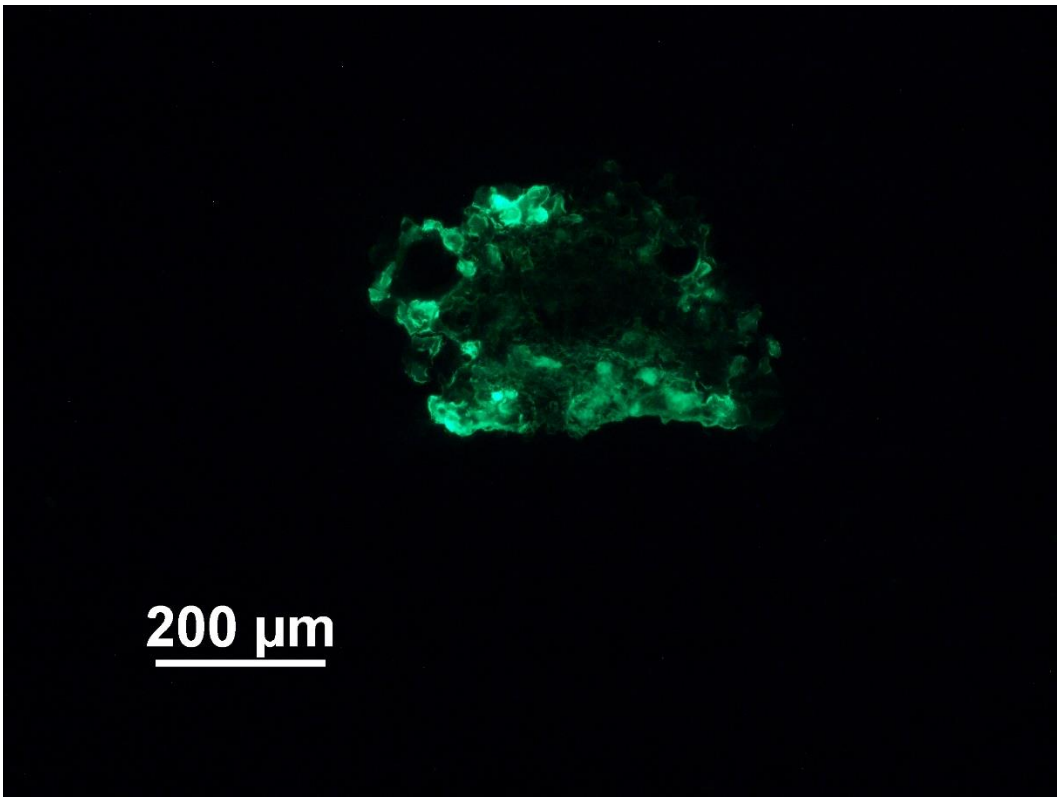

Figure 3-6 Fluorescence microscopy image of the same section

The image was taken from the same section of fixed co-culture spheroid seeded with GFP labeled, IR resistant HCT-116 cell line and wild type HCT-116 cell line. The image was taken with 1 second as exposure and 4.00x as gain.

Although the overall structure of fixed spheroids was not observed to change drastically before sectioning under the microscope, the morphology of each

individual cross section was found to be significantly changed with shrunk cells and cavity regions inside the spheroid. We believed that the integrity of the scaffold of the spheroids not observed to be changed was due to stacking of multiple layers of cells. When the cavity region overlapped with cells from another layer, the overall 3D looks of the spheroid seemed intact despite microcavities inside due to fixation. The comparison with fixed heart tissue sections validated our hypothesis.

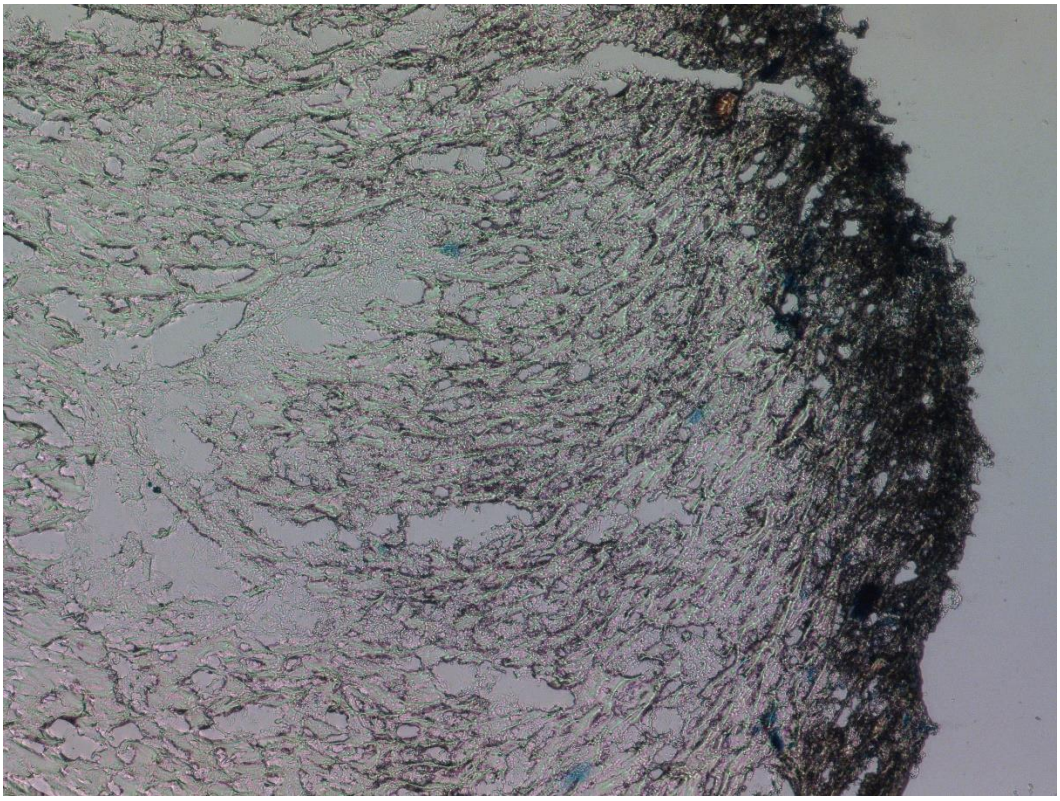

Figure 3-7 Bright field microscopy image of tissue from fixed heart

The shrinkage on the edge of the heart tissue was observed in other tissues as well, not only for spheroids. However, the change in the morphology was not obvious due to the low penetration of fixative on intact heart organ with larger size, but more significant for cancer cell spheroids that were smaller in size with gaps in between the cells forming the spheroids. Shapeshift due to fixation has also been reported in other literatures.[11] Compared with other MSI samples where the overall morphology was not significantly different despite fixation, other strategies for differentiating sources of cells in the co-cultured spheroids must be found.

3.5. Conclusion

For the samples mentioned above, different difficulties were met during the sample preparation process due to small sample size, unideal mechanical properties, or damage to the sample morphology, therefore the original plans of corresponding projects were paused.

3.6. References

(1) Shukla, U. V.; Tripathy, K. Diabetic Retinopathy. In *StatPearls*, StatPearls Publishing

Copyright © 2023, StatPearls Publishing LLC., 2023.
(2) Duh, E. J.; Sun, J. K.; Stitt, A. W. Diabetic retinopathy: current understanding, mechanisms, and treatment strategies. *JCI Insight* **2017**, *2* (14). DOI: 10.1172/jci.insight.93751.
(3) Lundeen, E. A.; Burke-Conte, Z.; Rein, D. B.; Wittenborn, J. S.; Saaddine, J.; Lee, A. Y.; Flaxman, A. D. Prevalence of Diabetic Retinopathy in the US in 2021. *JAMA Ophthalmology* **2023**, *141* (8), 747-754. DOI: 10.1001/jamaophthalmol.2023.2289 (acccessed 11/17/2023).
(4) Fruttiger, M. Development of the retinal vasculature. *Angiogenesis* **2007**, *10* (2), 77-88. DOI: 10.1007/s10456-007-9065-1.
(5) Łukasiewicz, S.; Czeczelewski, M.; Forma, A.; Baj, J.; Sitarz, R.; Stanisławek, A. Breast Cancer-Epidemiology, Risk Factors, Classification, Prognostic Markers, and Current Treatment Strategies-An Updated Review. *Cancers (Basel)* **2021**, *13* (17). DOI: 10.3390/cancers13174287.
(6) Park, M.; Kim, D.; Ko, S.; Kim, A.; Mo, K.; Yoon, H. Breast Cancer Metastasis: Mechanisms and Therapeutic Implications. *Int J Mol Sci* **2022**, *23* (12). DOI: 10.3390/ijms23126806.
(7) Hong, R.; Xu, B. Breast cancer: an up-to-date review and future perspectives. *Cancer Communications* **2022**, *42* (10), 913-936. DOI: https://doi.org/10.1002/cac2.12358.
(8) Białkowska, K.; Komorowski, P.; Bryszewska, M.; Miłowska, K. Spheroids as a Type of Three-Dimensional Cell Cultures-Examples of Methods of Preparation and the Most Important Application. *Int J Mol Sci* **2020**, *21* (17). DOI: 10.3390/ijms21176225.
(9) Xie, P.; Zhang, H.; Wu, P.; Chen, Y.; Cai, Z. Three-Dimensional Mass Spectrometry Imaging Reveals Distributions of Lipids and the Drug Metabolite Associated with the Enhanced Growth of Colon Cancer Cell Spheroids Treated with Triclosan. *Analytical Chemistry* **2022**, *94* (40), 13667-13675. DOI: 10.1021/acs.analchem.2c00768.
(10) Tian, X.; Zhang, G.; Zou, Z.; Yang, Z. Anticancer Drug Affects Metabolomic Profiles in Multicellular Spheroids: Studies Using Mass Spectrometry Imaging Combined with Machine Learning. *Analytical Chemistry* **2019**, *91*. DOI: 10.1021/acs.analchem.9b00026.
(11) Augusteyn, R. C.; Vrensen, G.; Willekens, B. The effect of paraformaldehyde fixation and PBS storage on the water content of the human lens. *Mol Vis* **2008**, *14*, 90-94.

Chapter 4 Study of Chemical Profile change of *Trypanosoma cruzi* Infected Heart after Fixation and Staining Using Single-Probe Mass Spectrometry

This is a collaborative project with Dr. Laura-Isobel McCall's group at the University of Oklahoma/San Diego State University. *Trypanosoma cruzi* infection, harvest and fixation of mice heart was conducted by Dr. McCall's group. The follow-up of this project has been performed by Dan Chen using DESI imaging with a manuscript under preparation.

4.1. Abstract

Chagas disease is a disease caused by the parasite *Trypanosoma cruzi* (*T. cruzi*), with possible cardiac and gastrointestinal complications during its chronic phase lasting lifelong.[1] Heart failure and cardiac arrest can occur as lethal danger to the patients with other lifelong risks if not treated properly. It has been estimated that 8 million people in the America region are having Chagas Disease.[2]

Fixation and staining are two commonly used methods for sample processing both adopted to study Chagas disease due to the parasitic residence, but their

compatibility with mass spectrometry (MS) is yet to be studied for a variety of reasons.[3, 4] Single-probe mass spectrometry was adopted to study the chemical profile change with and without fixation and/or staining treatment, in order to provide detail picture on how research could be performed with such sample treatment with minimum interference with mass spectrometry.

4.2. Introduction

Due to the wide compatibility and high sensitivity of mass spectrometry, MS has been broadly adopted in clinical and pharmaceutical applications for biomarker discovery, disease pathogenesis, and drug metabolism and pharmacokinetics.[5-8] Compared with vacuum-based MS methods which require more complicated sample preparation steps, ambient-based MS methods are gaining popularity for the simplified sample preparation steps thus possibility of analyzing biological samples at their native state.[9] Fixation and staining are two commonly used methods for sample processing.[3] For patient samples or infectious pathogens, fixation is a necessary step due to safety concerns. In the field of biological studies, staining is a popular technique for specific, intuitive spatial mapping of species of interest with the assistance from antibody, dyes, or fluorescence probes.[3, 10, 11] However, the compatibility between mass spectrometry and these methods was not studied and reported in detail in previous studies for a variety of reasons, especially when

unfixed patient samples were not available due to safety or degradation concerns. In this chapter, the effect of fixation and staining is studied.

4.3. Methods

4.3.1. Single-probe fabrication

The Single-probe was fabricated following the protocol established in previous publications.[12] Briefly summarized, the probe consisted of three parts including a dual bore quartz tip and two capillaries. The dual bore quartz tip was pulled from dual bore quartz tubing (O.D. 500 μm; I.D. 127 μm) with a laser puller. Two capillaries (O.D. 105 μm; I.D. 40 μm) were inserted into the two bores and fixed with UV glue. The assembled probe was fixed on a piece of glass slide for easier handling by epoxy glue.

4.3.2. Tissue fixation

Chemical fixation with glutaraldehyde was conducted on mice hearts provided by McCall group at University of Oklahoma.[13] Hearts taken from healthy mice were first pumped with 1x PBS to remove blood inside, and then immersed in glutaraldehyde solution for 15 minutes, followed by rinsing with 1x PBS.

4.3.3. Cryo-section

Fixed and unfixed heart samples were stored in -80 freezer prior to further processing. Embedding material was selected as 10% (m/v) Hydroxypropyl methylcellulose (HPMC) water solution for MS compatibility. The whole frozen hearts were taken out from the original container and completely immersed in room temperature embedding material for orientation adjustment. The mold containing immersed heart and embedding material was placed in dry ice for solidifying. The hardened material was sectioned into slices of 16 micrometer thickness and mounted onto microscope glass slides.

4.3.4. Staining

The staining protocol was adopted from previous publications.[14] To briefly summarize, the microscope slides were soaked with 0.2 mM MgCl2 in 1x PBS for 5 minutes for rehydration, then incubated with 2mM MgCl2, 4mM potassium ferricyanide, 4mM potassium ferrocyanide and X-gal at 2mg/ML in 1x PBS for 16-18 hours at 37 degrees Celsius. The stained sample was rinsed by 1x PBS after incubation to remove excess X-gal. For validation of the staining protocol, slides holding infected heart mice sections were included in every batch. All stained slides underwent fixation before staining.

4.3.5. Mass spectrometry

The Single-probe was held by a robotic arm to prevent vibration and slippery. Samples were placed and fixed onto an XYZ-translational stage controlled by LabView software and monitored with a digital microscope for precise location adjustment during the spot sampling and MSI process. Sampling solvent (e.g. 85%/15% methanol : water (v/v)) was provided continuously with a flow rate as 0.2 microliter per minute using a syringe pump. MS spectra was acquired using a Thermo LTQ Orbitrap XL mass spectrometer and LTQTune software.

4.4. Results and discussion

4.4.1. Microscopy and staining

Slides were examined under a microscope to verify the effectiveness of the staining process. As shown in the figures (Fig. 4-1, Fig. 4-2), blue spots indicating the residence of *Trypanosoma cruzi* on heart tissues only showed up on stained slides for infected mice, but not on stained slides for uninfected mice or unstained slides for infected mice, which validated the staining method for parasitic residence mapping.

Figure 4-1 Tissue section of infected mice heart after X-gal staining

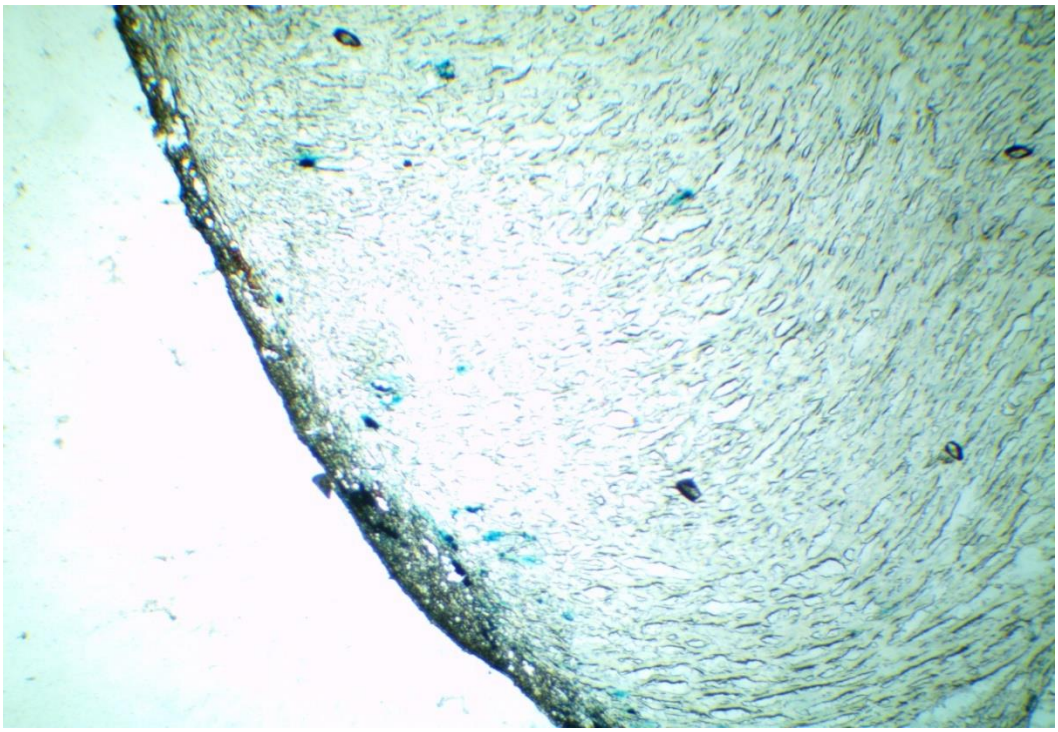

Figure 4-2 Zoomed view of the infected heart tissue section showing the stained parasitic spots

4.4.2. Mass spectrometry imaging

For feasibility study, mass spectrometry imaging was performed on infected, fixed heart tissue for investigation. Due to the lack of landscape structures in the myocardium region, only a rough comparison was made to locate the scanned region (Fig. 4-3).

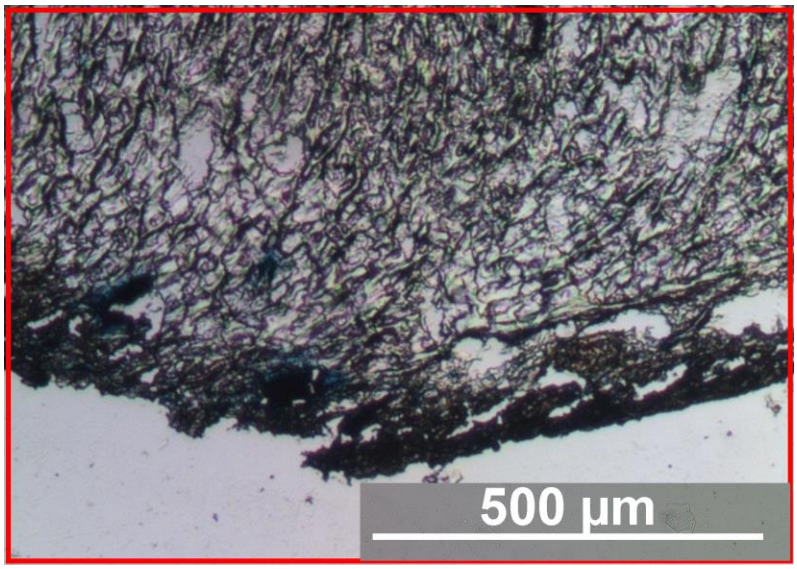

Figure 4-3 Microscope image of scanned myocardium tissue

From the collected MSI data, ion images for the corresponding region were constructed. Among all ion images, different patterns were observed with examples shown in Fig. 4-4. The ion image generated with total ion current (TIC) showed the boundary of the heart tissue, indicating that MSI could be a useful method for heart tissue studies. The detection of m/z 782.56 (tentatively labeled as PC(36:4) through database search), which was commonly observed in Single-probe SCMS studies for mammalian cells, also validated the feasibility of MSI method. Special features were found for other ions, with m/z 590.32 (tentatively labeled as LPC(22:6) through database search) and m/z 431.96 (no matches in current database within 10 ppm) as two examples.

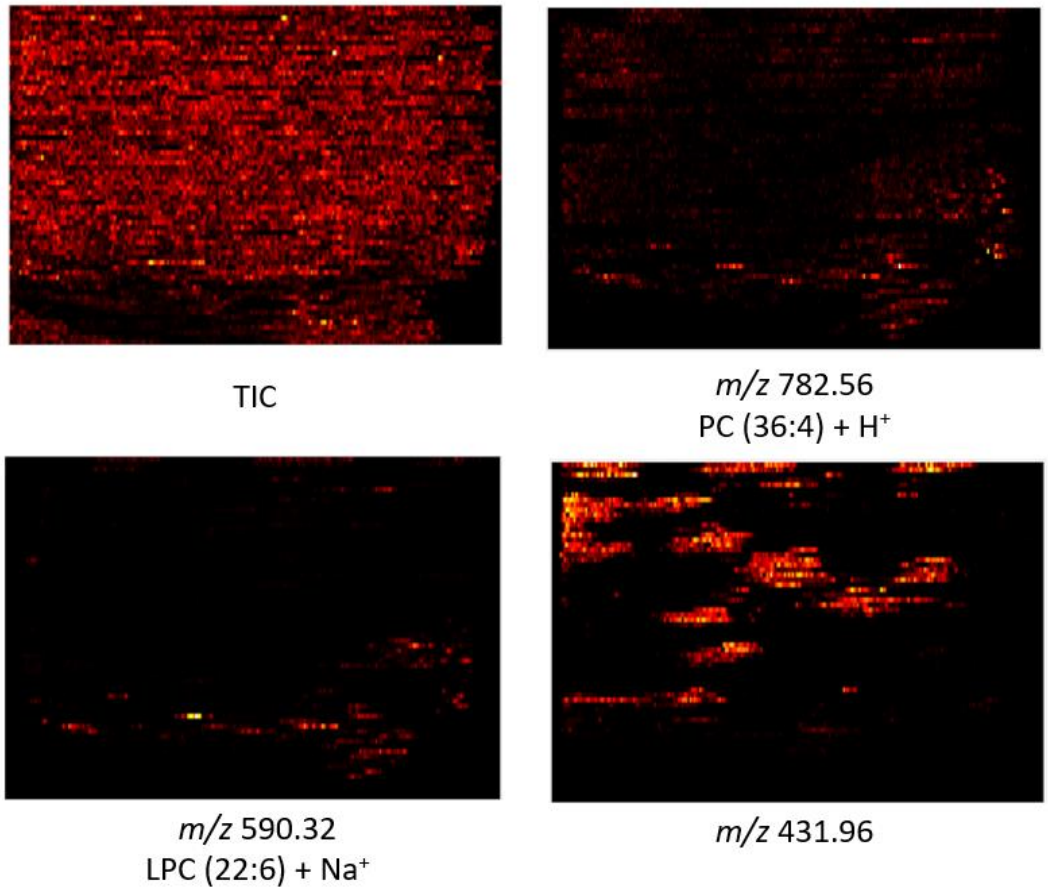

Figure 4-4 Ion images of fixed, infected mice heart tissue with tentative labels

In order to make connections between observed MSI features and parasitic infection spot of the mice heart, pre-MSI X-gal staining was purposed.

4.4.3. Investigation on pre-MSI staining

As the size of parasitic spots were observed to be similar or even smaller than the thickness of tissue sections, using adjacent slides for result comparison can lead to misinterpretation of data because the parasitic spots may only exist on one of the adjacent slides. In this case, the strategy to stain on the same tissue section must be applied. However, due to the possible delocalization of parasites on the tissue after our liquid extraction-based Single-probe MSI method, we tested the feasibility of pre-MSI staining to study the heart tissue.

To reduce the intrinsic biological variance among samples and batch effect due to shifted experimental conditions, consecutive slides from the same healthy heart was selected and treated differently on the same day for comparison (Table 4-1).

Table 4-1 Experimental setup for spot analysis of different treatment on mice myocardium tissues

Day 1		Day 2	
Group 1	untreated	Group 4	Washed then incubated in staining solution for 18h (regularly stained)
Group 2	untreated	Group 5	Washed then incubated in solvent of staining solution for 18h
Group 3	untreated	Group 6	Washed then stored in -80 freezer for 18h

MS data was collected from 10 spots on each heart in the myocardium region. The spectra obtained from six different groups was compared using analysis of variance (ANOVA) and principal component analysis (PCA).

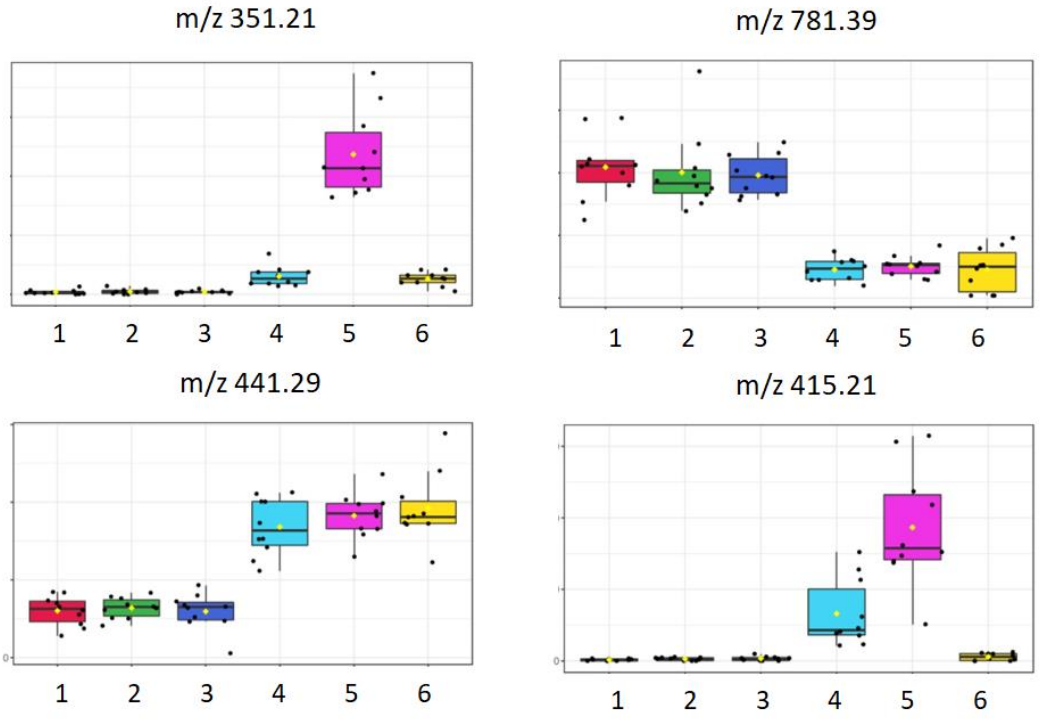

Figure 4-5 ANOVA result examples

From the ANOVA result (Fig. 4-5), untreated groups 1, 2, 3 were highly consistent with each other. However, different patterns for different ions were found among treated groups. For instance, abundances of m/z 781.39 and m/z 441.29 were simultaneously decreased and increased in group 4, 5, 6 compared with group 1, 2, 3, respectively, indicating that the abundance of those species might be sensitive to ambient environment exposure during the MS analysis process on the first day prior to any treatment. Significant differences were also found between group 4 and group 5 (e.g., m/z 351.21), showing that chemical species in the staining solution could possibly alter the chemical profile.

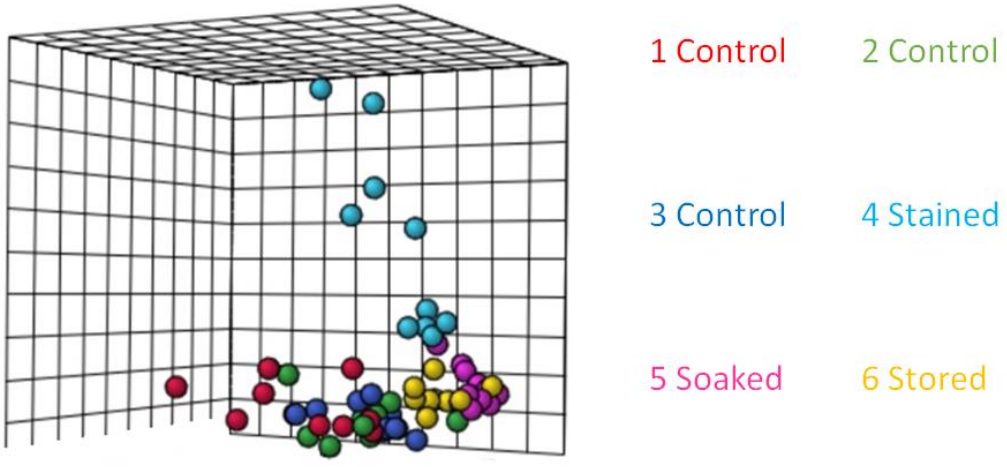

Figure 4-6 3D PCA result

Another perspective of view was obtained from PCA, which is a linear dimensionality reduction tool to study the global variance among high-dimensional observations. According to the 3D PCA result, stained group (cyan) was found to be well separated from other groups, indicating significant changes on the overall chemical profile of stained tissue.

The X-gal staining process has created a drastic change in the chemical profile of uninfected mice heart tissues detected with Single-probe MS method. The result suggested that using the same tissue section for MS analysis after staining operation could potentially lead to biased results. Although one of the popular strategies in MSI studies was using the adjacent section for comparison instead of the original

section, the method should not be applied to studies when the adjacent section of tissue is anticipated to have different morphology from the original section.

4.5. Conclusion

Unlike other studies where adjacent sections can be selected for comparison between results from different methods, the small spot size at several micrometers of parasitic residence blocks the possibility of using the same strategy. Although Single-probe MS results indicated that staining should not be performed before actual MS measurement, post-MS staining can still be performed for non-destructive MS sampling methods. This work has been carried on by Dan Chen using DESI imaging, which is a commercialized method with higher robustness.

4.6. References

(1) Lidani, K. C. F.; Andrade, F. A.; Bavia, L.; Damasceno, F. S.; Beltrame, M. H.; Messias-Reason, I. J.; Sandri, T. L. Chagas Disease: From Discovery to a Worldwide Health Problem. *Front Public Health* **2019**, *7*, 166. DOI: 10.3389/fpubh.2019.00166.
(2) *CDC - Chagas Disease*. Centers for Disease Control and Prevention, 2022. https://www.cdc.gov/parasites/chagas/gen_info/detailed.html#intro (accessed 2023 Nov 11).
(3) Buckner, F. S.; Wilson, A. J.; Van Voorhis, W. C. Detection of live Trypanosoma cruzi in tissues of infected mice by using histochemical stain for beta-galactosidase. *Infect Immun* **1999**, *67* (1), 403-409. DOI: 10.1128/iai.67.1.403-409.1999.
(4) Gordon, A.; Kannan, S. K.; Gousset, K. A Novel Cell Fixation Method that Greatly Enhances Protein Identification in Microproteomic Studies Using Laser Capture Microdissection and Mass Spectrometry. *Proteomics* **2018**, *18* (11), e1700294. DOI: 10.1002/pmic.201700294.
(5) Gowda, G. A.; Djukovic, D. Overview of mass spectrometry-based metabolomics: opportunities and challenges. *Methods Mol Biol* **2014**, *1198*, 3-12. DOI: 10.1007/978-1-4939-1258-2_1.
(6) Olshina, M. A.; Sharon, M. Mass Spectrometry: A Technique of Many Faces. *Q Rev Biophys* **2016**, *49*. DOI: 10.1017/s0033583516000160.
(7) Buchberger, A. R.; DeLaney, K.; Johnson, J.; Li, L. Mass spectrometry imaging: a review of emerging advancements and future insights. *Analytical chemistry* **2018**, *90* (1), 240.
(8) Zhang, L.; Vertes, A. Single-Cell Mass Spectrometry Approaches to Explore Cellular Heterogeneity. *Angew Chem Int Ed Engl* **2018**, *57* (17), 4466-4477. DOI: 10.1002/anie.201709719.
(9) Cahill, J. F.; Riba, J.; Kertesz, V. Rapid, untargeted chemical profiling of single cells in their native environment. *Analytical chemistry* **2019**, *91* (9), 6118-6126.
(10) Tian, X.; Xie, B.; Zou, Z.; Jiao, Y.; Lin, L.-E.; Chen, C.-L.; Hsu, C.-C.; Peng, J.; Yang, Z. Multimodal Imaging of Amyloid Plaques: Fusion of the Single-Probe Mass Spectrometry Image and Fluorescence Microscopy Image. *Analytical Chemistry* **2019**, *91* (20), 12882-12889. DOI: 10.1021/acs.analchem.9b02792.
(11) Larson, K.; Ho, H. H.; Anumolu, P. L.; Chen, M. T. Hematoxylin and Eosin Tissue Stain in Mohs Micrographic Surgery: A Review. *Dermatologic Surgery* **2011**, *37* (8).
(12) Rao, W.; Pan, N.; Yang, Z. High Resolution Tissue Imaging Using the Single-probe Mass Spectrometry under Ambient Conditions. *J Am Soc Mass Spectrom* **2015**, *26* (6), 986-993. DOI: 10.1007/s13361-015-1091-4.

(13) Nguyen, T. D.; Lan, Y.; Kane, S. S.; Haffner, J. J.; Liu, R.; McCall, L.-I.; Yang, Z. Single-Cell Mass Spectrometry Enables Insight into Heterogeneity in Infectious Disease. *Analytical Chemistry* **2022**, *94* (30), 10567-10572. DOI: 10.1021/acs.analchem.2c02279.
(14) Burn, S. F. Detection of β-galactosidase activity: X-gal staining. *Methods Mol Biol* **2012**, *886*, 241-250. DOI: 10.1007/978-1-61779-851-1_21.

Chapter 5 MassLite: An Integrated Python Platform for Single Cell Mass Spectrometry Metabolomics Data Pretreatment with Graphical User Interface and Advanced Peak Alignment Method

Part of the MS data for platform testing and troubleshooting was acquired by Zongkai Peng, Yunpeng Lan, Tra Nguyen, Shakya Wije Munige and Deepti Bhusal. The development of the platform was solely done by me with one manuscript under preparation.

5.1. Abstract

Mass spectrometry (MS) has been one of the most widely used tools for bioanalytical analysis due to its high sensitivity, capability of quantitative analysis, and compatibility with biomolecules. Among various MS techniques, single cell mass spectrometry (SCMS) is an advanced approach to molecular analysis of cellular contents in individual cells. In tandem with the creation of novel experimental techniques, the development of new SCMS data analysis tools is equally important. As most published software packages are not specifically designed to be compatible with improvised single cell sampling process, their

applicability on SCMS data is generally limited. Hereby we introduce a Python platform, MassLite, specifically designed for fast inhomogeneous SCMS metabolomics data pretreatment, with a peak alignment method that avoids binning in order to better process MS data obtained from high resolution mass spectrometers. The platform is made user-friendly with graphical user interface (GUI) and exports data in the forms of each individual cell for further analysis.

5.2. Introduction

Mass spectrometry (MS) has been playing an increasingly important role in the field of chemistry and bioanalysis since the invention of electrospray ionization (ESI)[1] and matrix-assisted laser desorption/ionization (MALDI)[2]. Assisted with improved sensitivity, resolution, and throughout of mass spectrometers[3] as well as advancement of computing power from hardware and software algorithm, MS has been broadly adopted in applications such as proteomics[4], metabolomics[5], biomarker discovery[6], and drug discovery.[7]

Among various experimental methods, liquid chromatography–MS (LC–MS) has especially been widely-applied, which excels in the separation and quantification of complex mixtures and biological samples.[8] However, due to the obligatory sample preparation, some critical information, such as spatial distribution of molecular species in tissues and cell heterogeneity, is inevitably lost from LC-MS

measurements. To overcome disadvantages of traditional LC-MS techniques, novel MS methods have been developed. Among them, MS imaging (MSI) is capable of offering insights into the spatial distribution of compounds in tissues, providing knowledge in histopathology and drug distribution.[9-11] In addition, single cell MS (SCMS) has recently gained increasingly more popularity due to its capability of reaching subcellular resolution and performing cell heterogeneity analysis.[12] Compared with traditional bulk analysis, SCMS reveals the chemical profiles of individual cells, providing unique understanding of complicated cell activities controlled by numerous intracellular and extracellular factors.

With the application of newly developed mass spectrometers possessing higher mass resolution, faster scan rate, and better sensitivity, the size of MS data has significantly increased in modern bioanalysis. Particularly, accurate mass measurements provide crucial information for molecular identification in metabolomics studies. Numerous software packages have been designed to pretreat the experiment data (i.e., peak picking, peak alignment, and intensity normalization) and extract essential information from data acquired from traditional LC-MS (e.g., MZmine[13-15] and XCMS[16, 17]) and novel MSI (e.g., Cardinal[18] and Metaspace[19]) experiments. In fact, some of these tools have been utilized to analyze certain types of SCMS data such as single cell proteomics, which require LC separation prior to MS analysis, and MALDI-based single cell metabolomics, which were acquired using strategies similar to MALDI techniques.[20-22] However, very few attempts

have been made to handle data acquired from ambient SCMS metabolomics.[23-26] Unlike most well-established LC-MS and MSI experiments, which are generally conducted using programmed, pre-loaded sampling and data acquisition process, most ambient SCMS metabolomics studies of live cells are commonly associated with improvised single cell sampling and segmented signal due to experimental conditions and operations, causing incompatibility with existing data processing tools that were designed for LC-MS and MSI.[27, 28] As any separation can potentially induce sample loss and dilution, in the SCMS studies of small molecules (e.g., metabolites), analyte separation is generally not included. Analysis of dense peaks heavily relies on the accurate mass detection from the spectra. Therefore, retaining all the valuable information obtained from high resolution mass spectra is a crucial need for SCMS metabolomics data pre-processing.

Among all data-processing steps (e.g., peak picking, peak alignment, intensity normalization), peak alignment is the step compensating for the slight m/z value variation of peaks from the same substance, so-called mass shift. The step of peak alignment corrects the mass shift and determines the actual m/z values of the peaks for further advanced analysis (e.g., multivariate analysis, data visualization, structure identification). Without alignment, signal from the same substance can be mistakenly split into multiple peaks, while a poor alignment might merge ions from different substance into same peaks, leading to a misinterpretation of MS data. Additional dimensions brought by improper peak alignment not only alter the

output of advanced data analysis methods, but also significantly increase the cost for further processing. To date, binning is the most widely used method for peak alignment. Briefly, for the convenience of data analysis, the entire range of the mass-to-charge ratio (m/z) of a mass spectrum is divided into a large number of equidistant small chunks, e.g. 0.1 Da, through a histogram-based method usually referred to as "binning".[29, 30] Although binning can significantly reduce the computational cost, this method possesses multiple intrinsic drawbacks.[30-32] First, the outcome of data processing is influenced by the parameters of bins, including bin width and bin position. Peaks could be artificially merged, split, or shifted due to unideal bin parameters, resulting in a loss of information. Second, using linear equidistant bins can lead to unequal mass error (i.e., ppm) of MS measurement. For example, as a commonly used bin width, 0.01 Da mass difference corresponds to 100 ppm at 100 Da, but 5 ppm at 2,000 Da. Third, binning cannot take full advantage of the capability of the high resolution of modern mass spectrometers. Although high resolution mass spectrometers can be used to provide accurate measurements (i.e., m/z values) of numerous ions, their molecular information, particularly for small molecules, such as metabolites, in complex samples cannot be efficiently extracted due to these above artifacts.

In this study, we introduce MassLite, a user-friendly, Python-based platform with graphical user interface (GUI) specifically designed for SCMS data pre-treatment. Compared with the existing SCMS metabolomics analysis tools, this new software

package possesses multiple advantages. First, this platform is robust to handle SCMS data acquired from intermittent acquisition processes, in which ion signals from individual cells are sequentially segmented. Second, MassLite can take full advantage of high-resolution mass spectrometer for detection of peaks with high mass accuracy. Third, automatic cell region selection is used to replace the existing labor-intensive manual process to increase processing throughput. Fourth, algorithm of peak alignment and background removal have been improved to be specifically compatible with SCMS metabolomics data. Last, the computational cost was significantly reduced with our purposed dynamic grouping method. Although the capability of this tool was demonstrated using the data generated from the Single-probe SCMS method, data produced from other SCMS techniques and platforms can be also processed with MassLite to increase the efficiency for data pretreatment with adjustment in the parameters for modification.[23]

5.3. Method

5.3.1. SCMS Experiment

Single cell mass spectrometry data was acquired with a method previously established by our group.[33, 34] To briefly summarize, the Single-probe, a miniatured sampling device with a tip size around 10 micrometers, was used to sample individual single cells under a microscope. In this study, a human colorectal

carcinoma cell line (HCT-116) and a cervical cancer cell line (HeLa) were used as models to generate Single-probe SCMS data for platform development. Cell content was extracted at the tip of the Single-probe via liquid junction, then delivered into a Thermo LTQ Orbitrap XL mass spectrometer with the inline nano-ESI emitter on the Single-Probe.

5.3.2. Data pretreatment

Data obtained from the experiment must undergo pretreatment for automation in further advanced analysis. The data pretreatment includes format conversion, algebraic transformation, void scan selection, cell scan selection, peak alignment, background peak removal and data exportation (Fig. 5-1).

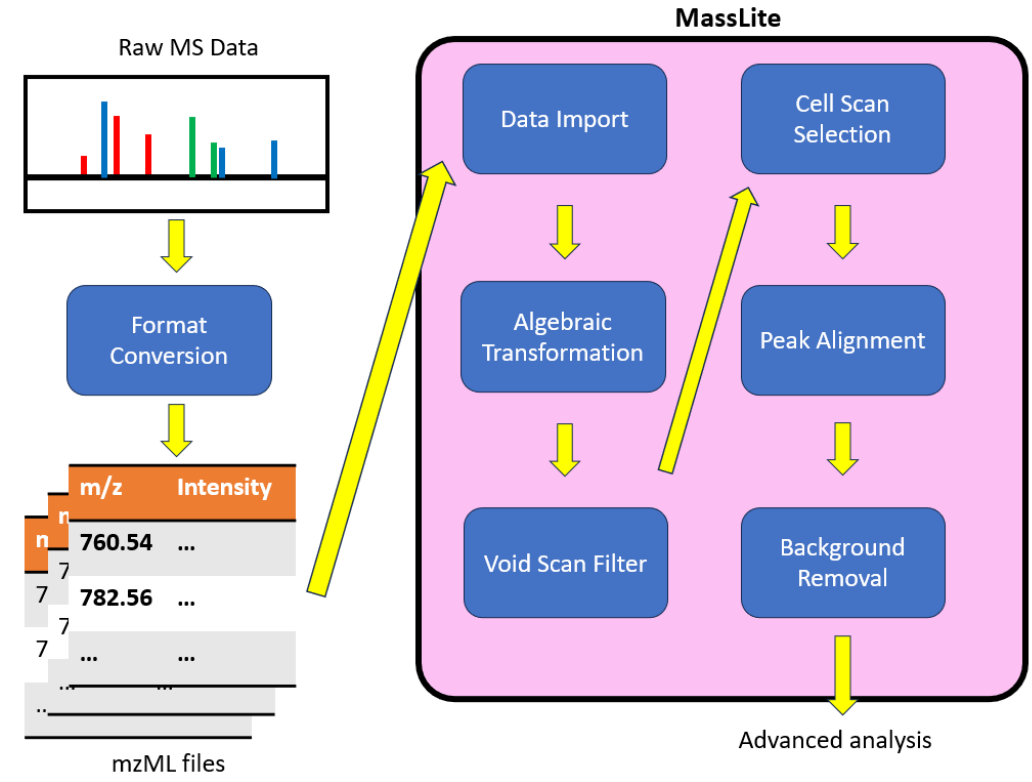

Figure 5-1 Schematic data processing workflow using MassLite

Except for the format conversion, which can be conducted using other existing tolls, the rest of the steps were all completed in MassLite. After data was imported into our platform, an additional algebraic transformation was first performed to use relative mass difference to describe difference among peaks instead of original linear difference. Next, a void scan filter was applied to distinguish intermittent scans during the data acquisition process. Then, the filtered scans were grouped by cells based upon the extracted ion chromatogram of selected cell markers.

Afterwards, peak alignment was performed with dynamic grouping to correct the mass shift of peaks. Last, background peaks can be selected to be removed along with other filters before exportation for further advanced analysis.

5.3.2.1. Data import

To make our platform compatible with SCMS metabolomics studies conducted with all types of mass spectrometers, the algorithm in MassLite was designed on basis of a universal MS data format, mzML.[35] For Single-probe SCMS data tested here, the original file generated from the Thermo Orbitrap LTQ XL mass spectrometer in .raw format, which was converted into the widely-used .mzML format using MSConvert. After format conversion, the data was read by our Python platform using pymzML package to extract the m/z values and intensities of the peaks. Peak picking was first conducted to obtain centroid peaks for each MS scan prior to further processing.

5.3.2.2. Algebraic transformation

As most MS studies utilize mass accuracy or mass measurement error (i.e., the difference between an individual measurement and the true value) in the unit of ppm, relative mass difference is a more stable and straightforward way to describe

the difference between two m/z values. In order to perform simpler peak matching during the pretreatment process, we performed a scaled, dynamic logarithmic transformation to intuitively describe the relative mass difference in the unit of ppm. In addition, this algorithm reflects mass accuracy with respect to m/z values, minimizing the influence of mass range on pick picking.

For two peaks at $\left(\frac{m}{z}\right)_1$ and $\left(\frac{m}{z}\right)_2$ (assuming $\left(\frac{m}{z}\right)_1 > \left(\frac{m}{z}\right)_2$), their relative mass difference can be described as:

$$\frac{\Delta ppm}{10^6} = [\left(\frac{m}{z}\right)_1 - \left(\frac{m}{z}\right)_2]/\{\frac{1}{2} \times \left[\left(\frac{m}{z}\right)_1 + \left(\frac{m}{z}\right)_2\right]\} \quad \text{(Eq. 1)}$$

When the peaks are very close to each other, the absolute difference is significantly smaller than their m/z values, thus $\left(\frac{m}{z}\right)_1 + \left(\frac{m}{z}\right)_2$ can be redeemed as $2 \times \left(\frac{m}{z}\right)_2$, we have

$$\frac{\Delta ppm}{10^6} = [\left(\frac{m}{z}\right)_1 - \left(\frac{m}{z}\right)_2]/\left(\frac{m}{z}\right)_2 \quad \text{(Eq. 2)}$$

By reorganizing the formula, we have

$$1 + \frac{\Delta ppm}{10^6} = \left(\frac{m}{z}\right)_1 / \left(\frac{m}{z}\right)_2 \quad \text{(Eq. 3)}$$

Taking logarithmic transformation on both sides, we have

$$\ln\left(1 + \frac{\Delta ppm}{10^6}\right) = ln\left(\frac{m}{z}\right)_1 - ln\left(\frac{m}{z}\right)_2 \quad \text{(Eq. 4)}$$

When two peaks are close enough to each other, Δppm is close to 0. Given that $\lim_{x\to 0} \ln(1+x) = x$ according to Taylor expansion, we have the following representation:

$$\frac{\Delta ppm}{10^6} = ln\left(\frac{m}{z}\right)_1 - ln\left(\frac{m}{z}\right)_2 \quad \text{(Eq. 5)}$$

Thus, when transformation $f\left(\frac{m}{z}\right) = \ln\left(\frac{m}{z}\right) \times 10^6$ was applied on two neighboring peaks, we have

$$f\left(\left(\frac{m}{z}\right)_1\right) - f\left(\left(\frac{m}{z}\right)_2\right) = [ln\left(\frac{m}{z}\right)_1 - ln\left(\frac{m}{z}\right)_2] \times 10^6 = \Delta ppm \qquad \text{(Eq. 6)}$$

Pairwise Euclidean distance between transformed m/z values can reflect the relative mass difference of the original m/z values in the unit of ppm, enabling fast processing and peak matching in the subsequent steps. In practical applications, to prevent non-negative values from being included in the calculations, a linear shift was included according to the lower limit of the mass range being detected.

5.3.2.3. Void scan removal

A typical ambient SCMS metabolomics dataset consists of informative scans (i.e., signals of cell analytes with coexisting solvent background and culture media) and void scans (i.e., scans containing only signals of instrument noise without identifiable species from cell analyte, solvent background, or culture media). The

void scans are usually inevitably included in data acquisition processes, mostly due to necessary operations during experiments such as relocation of the sampling or ionization device during the single cell isolation process. To automatically identify the void scans within the file, K-means, an unsupervised clustering method, was used to analyze intensity histogram of MS spectra for each scan. Because the intensity histogram reflects the general profiles of detected ions, significant changes of global pattern are expected between informative and void scan signal. For the actual K-means input, options of Uniform Manifold Approximation and Projection (UMAP) and logarithmic scaling are provided for transformation of the intensity histogram to enhance the discrimination between void scans and other scans. TIC (total ion current) of the clusters generated by the unsupervised K-means method can be visualized in the GUI for inspection, and clusters matching the definition of void scans can be dropped to reduce workload for the subsequent processes.

5.3.2.4. Cell and background region selection

To further increase the throughput of SCMS data processing, we developed an algorithm to automatically differentiate scans representing single cells from those from background such as solvent or cell culture media. First, a chromatogram was generated based upon the intensity of cell markers selected by users. For example,

m/z at 782.58 and 760.56 are commonly detected in cells, and they were selected as default indicators of single cell detection (i.e., marker signals). Second, an initial Gaussian smoothing was performed for extracted ion chromatogram (EIC) to avoid unideal splits for signals from each single cell due to signal fluctuation during the data acquisition process. Third, scans of cells and background were defined. After the local maxima and minima were primarily found, a finer global search resistant to signal fluctuation across the whole chromatogram was conducted. A stricter intensity requirement for peak search within maxima found in the previous search was applied to account for possible peak splitting issue due to signal fluctuation in the EIC. In the current study, the regions containing marker signals ≥20% of the local maxima were defined as cell regions, whereas region containing marker signals <5% of the local maxima were regarded as background regions unless otherwise defined.

5.3.2.5. Peak alignment with dynamic grouping

Due to multiple factors (e.g., the intrinsic performance of instrument and fluctuation of ion signals and instrument conditions), mass shift generally occurs during MS analysis.[36-38] Because accurate m/z values provide important information for molecular identification, mass shift is a critical error in high resolution MS studies, in which multiple ions with similar m/z values can be

simultaneously detected. Inappropriate handling of mass shift may result in artifacts such as peak splitting, loss of peaks, or inaccurate m/z assignment. To compensate for the mass shift across different scans, peak alignment must be performed. For the ease of processing and precise m/z value description, centroiding on all peaks was performed, keeping only one m/z value of peak center and one intensity value for each peak. All centroid peaks along with their transformed m/z values from all imported scans were included for peak alignment. Hierarchical clustering was performed for observed peaks to find internal matching among themselves. The desired mass shift tolerance was set as the cluster size for hierarchical clustering so that coverage for each aligned peak does not exceed the threshold. Thus, the mass accuracy of peak alignment was guaranteed. With the algebraic transformation performed in earlier steps, simple one-dimensional Euclidean distance can be redeemed as the relative mass difference between the m/z values of different peaks. To reduce the cost of pairwise distance calculation in the hierarchical clustering process, a “divide and conquer” strategy, so-called “dynamic grouping”, has been applied to split the data in chunks, eliminating unnecessary comparison of peaks from different data chunks which accounted for most of the cost from direct comparison (Fig. 5-2). To address potential peak splitting issues due to this strategy, a boundary checker was included to validate the relationship between two closest peaks in adjacent chunks. This binning-free method can maintain higher mass resolution from the original data.

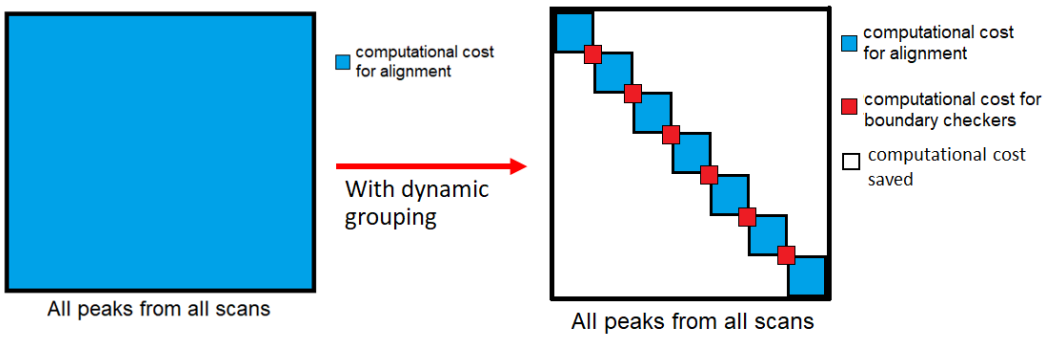

Figure 5-2 Schematic mechanism of dynamic grouping

5.3.2.6. Background removal

During ambient SCMS measurement, particularly for live cell analysis, interfering ions from impurities in solvent or species in cell culture media are generally detected along with cellular contents. To eliminate artifacts in analysis, interfering ions should be treated as background and excluded. Thus, aligned peaks with their highest intensities in one of the background scan regions, which could be determined automatically in the cell scan selection step, were regarded as the background substance and subsequentially filtered from the data. Compared with the traditional binning method for background removal, the algorithm does not require prior knowledge of the cell systems, and has the ability to distinguish peaks from background substances and cell analytes with highly similar m/z values to them.

5.4. Result and discussion

5.4.1. Graphical user interface

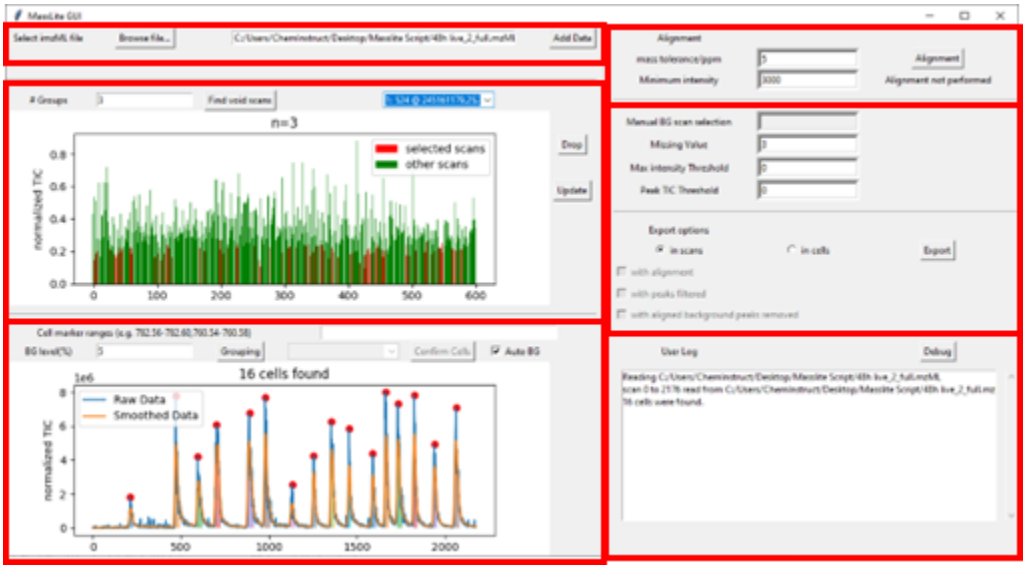

Figure 5-3 MassLite graphical user interface

The graphical user interface (GUI) was built using tkinter package in Python with functions run by packages, including numpy, scipy, matplotlib, and pandas. In our current design, the GUI has 6 major parts, including data read-in, void scan filter, cell sorting, peak alignment, exportation filter, and debugging modules (Fig. 5-3).

5.4.2. Parameter optimization for void scan filter

Because the global spectral features of the void scans are significantly different from those of informative scans containing analytes from cells or solvent background, the clustering was based on the intensity histogram, which mainly describes the overall structure of the whole MS spectrum. Different parameters for the generation of the intensity histogram were tested, along with two different techniques (e.g., logarithmic scaling, UMAP dimensionality reduction) aiming at enhancement of discrimination between void scans and other scans.

In our experience of handling SCMS metabolomics data of mammalian cells acquired using the Single-Probe SCMS techniques, lipid signals were significantly increased when cellular contents were extracted and detected, especially in the range of m/z 700-800 Da.[33, 39] On the other hand, ionization of cell analytes caused the suppression of the base peak in background scan, usually in the range of m/z 350-550 Da. This trend is also expected in studies using other SCMS platforms. To guarantee the features in the MS spectra, which are induced by the two major sources of signals mentioned above, can be captured in the intensity histogram, we tested both 50-Da and 100-Da intervals for histogram generation of MS1 data in the mass range of m/z 50-2000 Da. Smaller intervals for the histogram can maintain more details in the spectra structure, but also led to oversized machine learning input, thus deviating from our purpose for quick detection. On the other hand, larger

intervals could possibly fail to capture the changes in the spectra features if the increase and decrease of ion intensity occur within the same interval.

To perform clustering for efficient identification of void scans, different strategies to enhance discrimination between void scans and other scans have been applied. Based upon previous observations in Single-probe SCMS data, void scans usually contain lower signal intensities compared with other informative scans, so logarithmic scaling could be a quick, feasible strategy to isolate void scans with such property. Original intensity can more intuitively reflect the instant condition when the scan took place, while logarithmic transformations significantly reduce the difference of signal intensities and enhance the detection of low intensity ions. However, for experiments where decreased intensities for void scans are not observed, dimensionality reduction tools provide alternative options. UMAP, a powerful technique with relatively low computational cost compared with other nonlinear dimensionality reduction methods, has been adopted as an example and tested. The effect of logarithmic scaling and dimensionality reduction using UMAP were tested both individually and jointly.

Although data is not labelled beforehand for unsupervised clustering, certain criteria must be defined to match the goal of quickly identifying void scans through clustering. Given that the variance between scans of the same type of signal can hardly be estimated due to the heterogeneity among individual cells, the total

number of clusters would be a more practical parameter to guide the process compared with cluster variation. Based upon previous Single-probe SCMS observations, the possible major sources of each individual scan are cell analytes, solvent background, and cell culture media. With possible subpopulations existing within each type, a total number of cluster n≥3 (i.e., 3, 4, 5) would be a reasonable blind guess universal for different SCMS experiments.

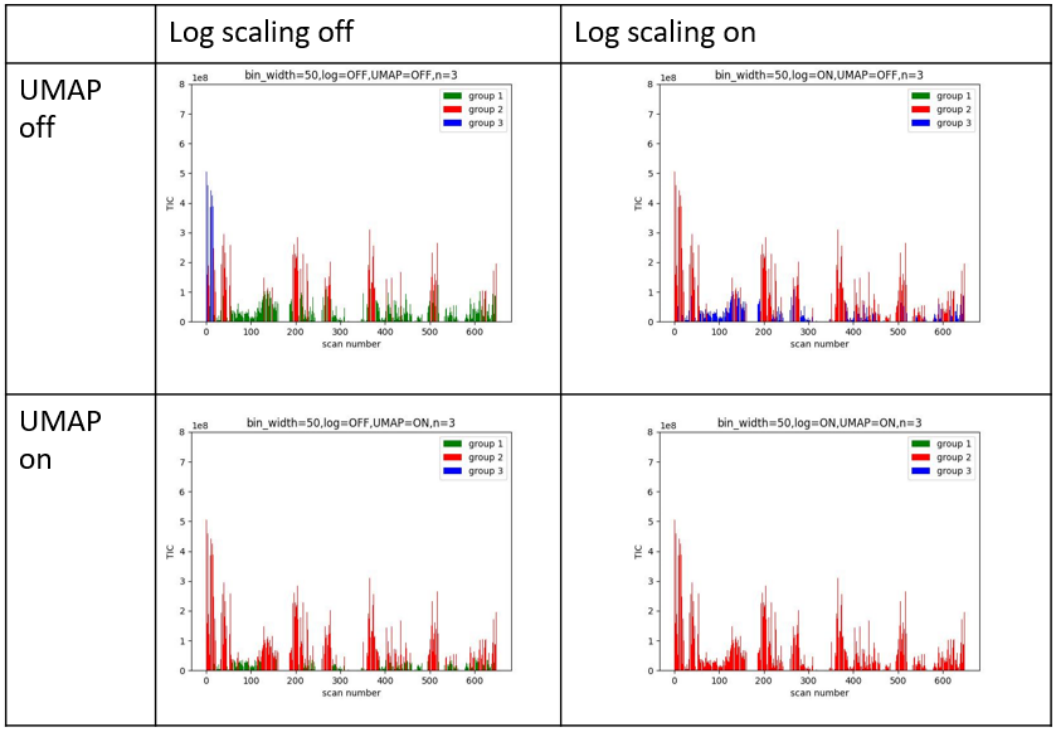

Figure 5-4 Examples of void scan filter parameter testing

We used a total of 24 combinations of different parameters mentioned above to test the same dataset (part of the results shown in Fig. 5-4). Different intervals for histogram generation seemed to be the least sensitive parameter because either 100 or 50 Da interval over m/z 100 – 2000 range provided enough features for the K-means clustering.

When untransformed data was used as the K-means input, the difference between void scans and low intensity scans was much less significant compared with signal fluctuation among high intensity scans in the untransformed intensity histogram, leading to insufficient discrimination between void scans and other types of scans in the clustering result. To address this issue, logarithmic and UMAP transformation were utilized to enhance the separation of scans with low intensities. Logarithmic scaling was adopted for void scans, which possess considerably lower ion signals compared with other informative scans. Alternatively, nonlinear dimensionality reduction can catch the similarity between each group of scans thus differentiate between void scans and other scans, and therefore UMAP was adopted as an example of nonlinear dimensionality reduction to treat the data. Both logarithmic and UMAP transformation provided satisfactory clustering output for the purpose of identifying void scans when working individually, with slight differences in the unimportant clusters that were not considered to be void. In contrast, satisfactory performance was not achieved using untransformed data, which contain low intensities of informative scans and void scans in the same

cluster. Although both transformations provide better classification outcomes than non-transformed data when working individually, simultaneously implementing both logarithmic scaling and UMAP tends to lead to undesired result, with unbalanced group sizes when the number of clusters n>3. To effectively sort out void scans, either logarithmic or UMAP transformation is thus adequate without causing artificial split and demanding substantial computing power.

5.4.3. Alignment result

To investigate the quality of our peak alignment algorithm, the alignment output of a 2176 scan dataset from MassLite was compared with those obtained from mspalign function in Matlab, which is a common programming language, as well as Geena 2 and MZmine 3, which are two widely used data processing platforms. Compared with mspalign function, MassLite produced peaks with diverse distribution and clearer background (Fig. 5-5). Equidistant aligned peaks are common artifacts due to the binning step in histogram-based methods, which cannot reflect the true nature of original MS spectra. Additional computational cost, which was induced by relatively high noise through the whole m/z axis, in further analysis was also found in the alignment result using mspalign.

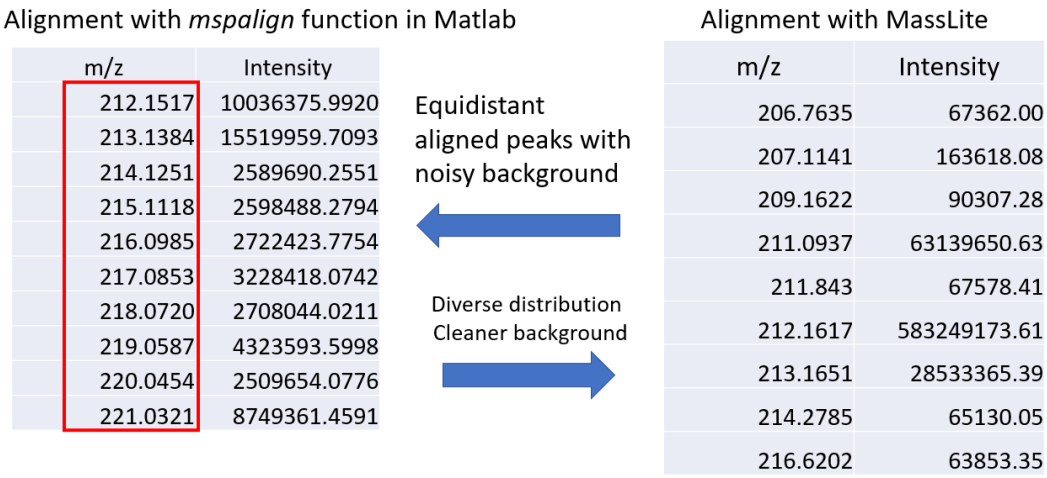

Figure 5-5 Comparison between mspalign and MassLite result

As for Geena 2, this online platform could not handle the whole dataset collected from 16 cells, which consist of 2176 scans with 5 ppm mass shift tolerance, likely due to the large size of the data. Alternatively, a truncated version of 4 cells from the first 830 scans at default 0.1 Da mass shift tolerance was submitted and processed by Geena 2. Both MZmine 3 and MassLite handled the original 2176 scan dataset. To investigate the mass accuracy maintained by each platform, all aligned peaks were re-ordered in ascending order and the relative mass difference between adjacent peaks were calculated using our algebraic transformation. The relative mass difference between adjacent peaks can reflect the ability of data processing platforms on resolving peaks with similar m/z values. For an intuitive

view, a histogram showing the distribution of relative mass differences between adjacent peaks was generated (Fig. 5-6, 5-7, 5-8).

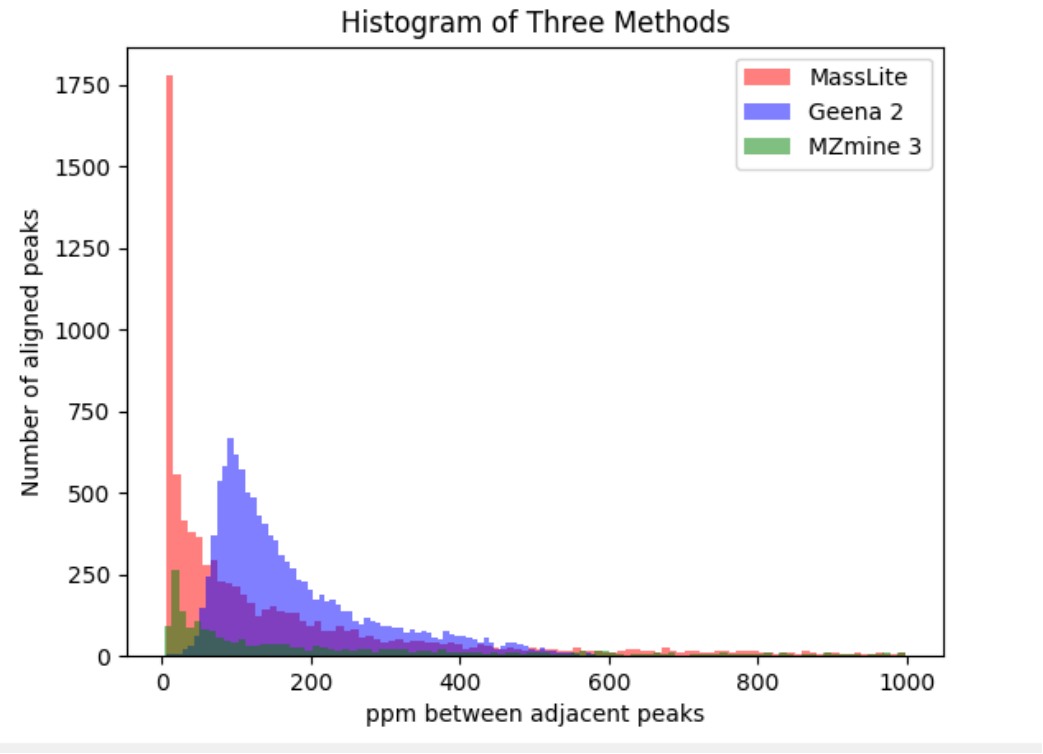

Figure 5-6 Histogram showing relative mass differences distribution of aligned peaks by MassLite, Geena 2, and MZmine 3

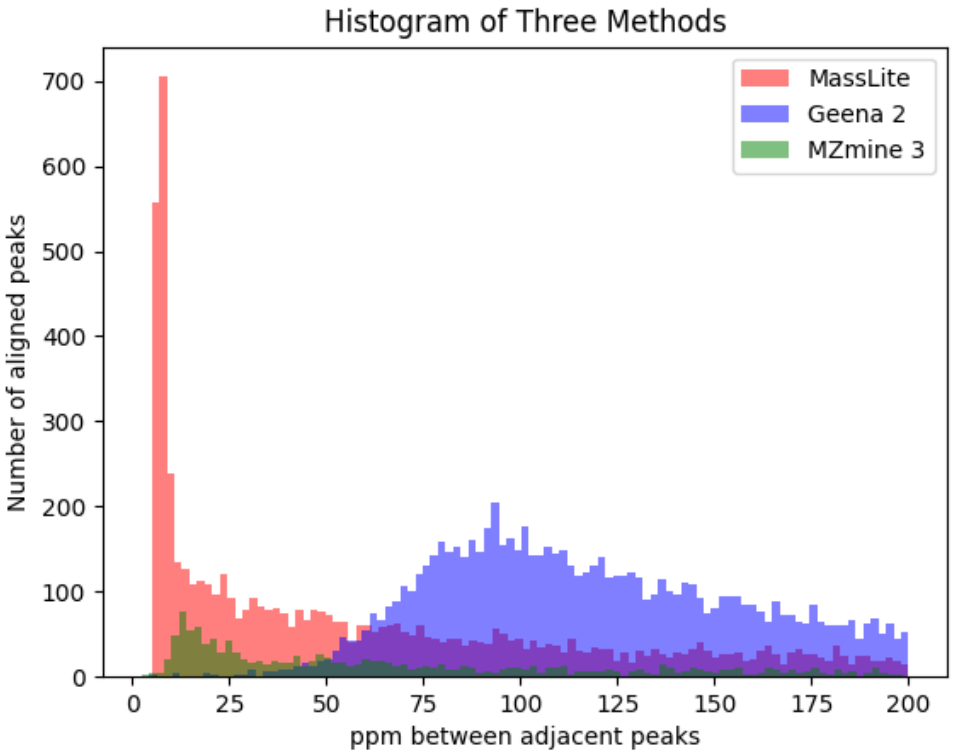

Figure 5-7 Zoomed view of the histogram comparing MassLite, Geena 2, and MZmine 3

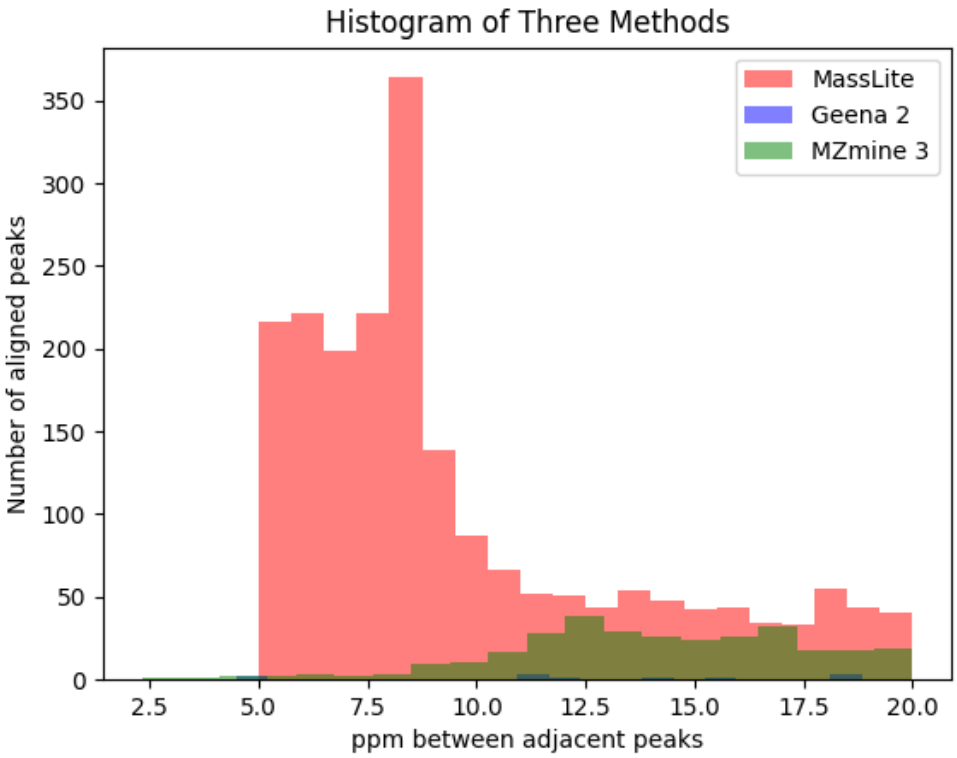

Figure 5-8 Zoomed view of the histogram at high resolution range (below 20 ppm)

From the histogram and its zoomed views, MassLite was able to differentiate more signals that were 5-10 ppm apart from each other, which showed its capability of aligning peaks from MS spectra at a higher resolution compared to two popular tools in the MS field. In fact, the 5-ppm cut-off was used in the current study, whereas users can determine the suitable values according to the specific studies.

Lower cut-off values can be potentially used to treat MS data acquired using mass spectrometers with higher resolving power.

5.4.4. Computational cost

Computational cost, including both CPU time and memory usage, for peak alignment was discussed for conditions with and without dynamic grouping. The computational cost depends on both the total number of scans and the total number of peaks in each scan. Because of the variance of the molecular composition in each particular scan, the number of peaks in each scan is subject to change y, thus the total number of peaks is positively correlated with number of scans but in a non-linear fashion. Among all data pretreatment steps, peak alignment without binning is the most expensive part due to the pairwise distance calculation and distance matrix update during hierarchical clustering. Regular pairwise comparison between peaks requires computational cost to the second power of total peaks detected. Although binning can reduce the index of power to one by replacing pairwise comparison among observed peaks with direct comparison between observed peaks and preset bins, the loss of mass accuracy during alignment step limited its applicability on SCMS metabolomics data, thus we purposed a dynamic grouping method (Fig. 5-2). Dynamic grouping reduces the computational cost for peak alignment with a similar ‘divide and conquer’ strategy. When the whole dataset was

divided into multiple chunks, the number of unnecessary comparisons between peaks was largely reduced. Particularly, comparisons between peaks from different chunks were eliminated, which could theoretically reduce the cost by second power to the number of chunks. To overcome the potential peak splitting issue due to the boundaries of the chunks, we implemented an automatic check at the boundary of adjacent neighboring chunks to merge artificially split peaks due to the chunk division, which slightly increased the cost by the first power to the number of chunks. As the actual cost reduction of dynamic binning depended on both dataset size, total number of peaks, total number of scans and number of chunks, we tested datasets with a small (16.2 MB imzML file, Table 5-1) and a large (2.83 GB imzML file, Table 5-2) sizes, finding that the time cost for peak alignment with dynamic grouping has been reduced to 1/3 and 1/10 at their ideal group sizes for grouping compared with direct pairwise comparison, respectively. In addition to time cost, memory usage is another major concern because storing all pairwise distances for millions of peaks, which lead to trillions of distances, can occupy several TB of memory, potentially resulting in breakdown of the algorithm. Dynamic grouping significantly reduced both CPU time and memory usage while providing reasonable results, allowing for customizable studies using a local computer. The method can be potentially improved when multiple cores are available for parallel processing.

Table 5-1 Time cost for peak alignment*

Group size using	2 x n	3 x n	10 x n	Without Grouping
Alignment cost (sec.)	3.17	3.13	4.53	9.21
Normalized cost	34%	34%	49%	100%

*Data size is 16.2 MB.

n: the group size, which is given in the unit of number of total scans in the file.

Table 5-2 Time cost for peak alignment*

Group size	1.5 x n	2 x n	3 x n	Without Grouping
Alignment cost (h)	5.4	7.1	10.4	>48
Normalized cost	11%	15%	22%	100%

*Data size is 2.83 GB.

n: the group size, which is given in the unit of number of total scans in the file.

5.5. Conclusion

In this paper we have introduced a Python-based platform for mass spectrometry data pretreatment, specifically designed for SCMS. Compared with previously established methods designed for LC-MS, MassLite enhances the mass resolution of detected peaks, can be run at ease with our graphical user interface, and provides real-time feedback from the output on a local computer. Especially for untargeted single-cell metabolomics without chromatographic separation, our method can prevent low-intensity peaks from being filtered, thus increasing the possibility of finding uncommonly seen substances among MS data. As all results were maintained in the application prior to exportation, the trade-off between ‘keeping more low-abundance signal’ and ‘removing more noise’ can be tuned by users easily with local GUI. Although the algorithm has only been applied to MS data collected from Thermo Orbitraps, we are optimistic about its applicability on MS data collected from other types of instruments.

5.6. References

(1) Fenn, J. B.; Mann, M.; Meng, C. K.; Wong, S. F.; Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* **1989**, *246* (4926), 64-71. DOI: 10.1126/science.2675315.
(2) Karas, M.; Hillenkamp, F. Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons. *Analytical chemistry* **1988**, *60* (20), 2299-2301.
(3) Crutchfield, C. A.; Clarke, W. Present and Future Applications of High Resolution Mass Spectrometry in the Clinic. *Discoveries (Craiova)* **2014**, *2* (2), e17. DOI: 10.15190/d.2014.9.
(4) Baxi, A. B.; Pade, L. R.; Nemes, P. Mass spectrometry based proteomics for developmental neurobiology in the amphibian Xenopus laevis. *Curr Top Dev Biol* **2021**, *145*, 205-231. DOI: 10.1016/bs.ctdb.2021.04.002.
(5) Vangeenderhuysen, P.; Van Arnhem, J.; Pomian, B.; De Graeve, M.; De Commer, L.; Falony, G.; Raes, J.; Zhernakova, A.; Fu, J.; Hemeryck, L. Y.; et al. Dual UHPLC-HRMS Metabolomics and Lipidomics and Automated Data Processing Workflow for Comprehensive High-Throughput Gut Phenotyping. *Analytical Chemistry* **2023**, *95* (22), 8461-8468. DOI: 10.1021/acs.analchem.2c05371.
(6) Martín-Blázquez, A.; Díaz, C.; González-Flores, E.; Franco-Rivas, D.; Jiménez-Luna, C.; Melguizo, C.; Prados, J.; Genilloud, O.; Vicente, F.; Caba, O.; et al. Untargeted LC-HRMS-based metabolomics to identify novel biomarkers of metastatic colorectal cancer. *Scientific Reports* **2019**, *9* (1), 20198. DOI: 10.1038/s41598-019-55952-8.
(7) Richter, L. H. J.; Jacobs, C. M.; Mahfoud, F.; Kindermann, I.; Böhm, M.; Meyer, M. R. Development and application of a LC-HRMS/MS method for analyzing antihypertensive drugs in oral fluid for monitoring drug adherence. *Anal Chim Acta* **2019**, *1070*, 69-79. DOI: 10.1016/j.aca.2019.04.026.
(8) Pitt, J. J. Principles and applications of liquid chromatography-mass spectrometry in clinical biochemistry. *Clin Biochem Rev* **2009**, *30* (1), 19-34.
(9) Paine, M. R. L.; Liu, J.; Huang, D.; Ellis, S. R.; Trede, D.; Kobarg, J. H.; Heeren, R. M. A.; Fernández, F. M.; MacDonald, T. J. Three-Dimensional Mass Spectrometry Imaging Identifies Lipid Markers of Medulloblastoma Metastasis. *Scientific Reports* **2019**, *9* (1), 2205. DOI: 10.1038/s41598-018-38257-0.
(10) Tian, X.; Xie, B.; Zou, Z.; Jiao, Y.; Lin, L.-E.; Chen, C.-L.; Hsu, C.-C.; Peng, J.; Yang, Z. Multimodal Imaging of Amyloid Plaques: Fusion of the Single-Probe Mass Spectrometry Image and Fluorescence Microscopy Image. *Analytical Chemistry* **2019**, *91* (20), 12882-12889. DOI: 10.1021/acs.analchem.9b02792.
(11) Xie, P.; Zhang, H.; Wu, P.; Chen, Y.; Cai, Z. Three-Dimensional Mass Spectrometry Imaging Reveals Distributions of Lipids and the Drug Metabolite

Associated with the Enhanced Growth of Colon Cancer Cell Spheroids Treated with Triclosan. *Analytical Chemistry* **2022**, *94* (40), 13667-13675. DOI: 10.1021/acs.analchem.2c00768.
(12) Zhang, L.; Vertes, A. Single-Cell Mass Spectrometry Approaches to Explore Cellular Heterogeneity. *Angew Chem Int Ed Engl* **2018**, *57* (17), 4466-4477. DOI: 10.1002/anie.201709719.
(13) Katajamaa, M.; Miettinen, J.; Orešič, M. MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics* **2006**, *22* (5), 634-636. DOI: 10.1093/bioinformatics/btk039 (acccessed 10/17/2023).
(14) Pluskal, T.; Castillo, S.; Villar-Briones, A.; Orešič, M. MZmine 2: Modular framework for processing, visualizing, and analyzing mass spectrometry-based molecular profile data. *BMC Bioinformatics* **2010**, *11* (1), 395. DOI: 10.1186/1471-2105-11-395.
(15) Schmid, R.; Heuckeroth, S.; Korf, A.; Smirnov, A.; Myers, O.; Dyrlund, T. S.; Bushuiev, R.; Murray, K. J.; Hoffmann, N.; Lu, M.; et al. Integrative analysis of multimodal mass spectrometry data in MZmine 3. *Nature Biotechnology* **2023**, *41* (4), 447-449. DOI: 10.1038/s41587-023-01690-2.
(16) Smith, C. A.; Want, E. J.; O'Maille, G.; Abagyan, R.; Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* **2006**, *78* (3), 779-787. DOI: 10.1021/ac051437y.
(17) Forsberg, E. M.; Huan, T.; Rinehart, D.; Benton, H. P.; Warth, B.; Hilmers, B.; Siuzdak, G. Data processing, multi-omic pathway mapping, and metabolite activity analysis using XCMS Online. *Nature Protocols* **2018**, *13* (4), 633-651. DOI: 10.1038/nprot.2017.151.
(18) Bemis, K. D.; Harry, A.; Eberlin, L. S.; Ferreira, C.; van de Ven, S. M.; Mallick, P.; Stolowitz, M.; Vitek, O. Cardinal: an R package for statistical analysis of mass spectrometry-based imaging experiments. *Bioinformatics* **2015**, *31* (14), 2418-2420. DOI: 10.1093/bioinformatics/btv146 (acccessed 10/18/2023).
(19) Palmer, A.; Phapale, P.; Chernyavsky, I.; Lavigne, R.; Fay, D.; Tarasov, A.; Kovalev, V.; Fuchser, J.; Nikolenko, S.; Pineau, C.; et al. FDR-controlled metabolite annotation for high-resolution imaging mass spectrometry. *Nature Methods* **2017**, *14* (1), 57-60. DOI: 10.1038/nmeth.4072.
(20) Lombard-Banek, C.; Moody, S. A.; Nemes, P. Single-Cell Mass Spectrometry for Discovery Proteomics: Quantifying Translational Cell Heterogeneity in the 16-Cell Frog (Xenopus) Embryo. *Angewandte Chemie International Edition* **2016**, *55* (7), 2454-2458. DOI: https://doi.org/10.1002/anie.201510411 (acccessed 2023/12/07).
(21) Virant-Klun, I.; Leicht, S.; Hughes, C.; Krijgsveld, J. Identification of Maturation-Specific Proteins by Single-Cell Proteomics of Human Oocytes.

Molecular & Cellular Proteomics **2016**, *15* (8), 2616-2627. DOI: https://doi.org/10.1074/mcp.M115.056887.
(22) Neumann, E. K.; Comi, T. J.; Rubakhin, S. S.; Sweedler, J. V. Lipid Heterogeneity between Astrocytes and Neurons Revealed by Single-Cell MALDI-MS Combined with Immunocytochemical Classification. *Angewandte Chemie International Edition* **2019**, *58* (18), 5910-5914. DOI: https://doi.org/10.1002/anie.201812892 (acccessed 2023/12/07).
(23) Shao, Y.; Zhou, Y.; Liu, Y.; Zhang, W.; Zhu, G.; Zhao, Y.; Zhang, Q.; Yao, H.; Zhao, H.; Guo, G.; et al. Intact living-cell electrolaunching ionization mass spectrometry for single-cell metabolomics. *Chemical Science* **2022**, *13* (27), 8065-8073, 10.1039/D2SC02569H. DOI: 10.1039/D2SC02569H.
(24) Liu, R.; Zhang, G.; Yang, Z. Towards rapid prediction of drug-resistant cancer cell phenotypes: single cell mass spectrometry combined with machine learning. *Chemical Communications* **2019**, *55* (5), 616-619, 10.1039/C8CC08296K. DOI: 10.1039/C8CC08296K.
(25) Liu, R.; Yang, Z. Single cell metabolomics using mass spectrometry: Techniques and data analysis. *Analytica Chimica Acta* **2021**, *1143*, 124-134. DOI: https://doi.org/10.1016/j.aca.2020.11.020.
(26) Liu, R.; Zhang, G.; Sun, M.; Pan, X.; Yang, Z. Integrating a generalized data analysis workflow with the Single-probe mass spectrometry experiment for single cell metabolomics. *Analytica chimica acta* **2019**, *1064*, 71-79.
(27) Pan, N.; Rao, W.; Yang, Z. Single-Probe Mass Spectrometry Analysis of Metabolites in Single Cells. *Methods Mol Biol* **2020**, *2064*, 61-71. DOI: 10.1007/978-1-4939-9831-9_5.
(28) Gholipour, Y.; Erra-Balsells, R.; Hiraoka, K.; Nonami, H. Living cell manipulation, manageable sampling, and shotgun picoliter electrospray mass spectrometry for profiling metabolites. *Anal Biochem* **2013**, *433* (1), 70-78. DOI: 10.1016/j.ab.2012.10.001.
(29) Finch, J. P.; Wilson, T.; Lyons, L.; Phillips, H.; Beckmann, M.; Draper, J. Spectral binning as an approach to post-acquisition processing of high resolution FIE-MS metabolome fingerprinting data. *Metabolomics* **2022**, *18* (8), 64. DOI: 10.1007/s11306-022-01923-6.
(30) Feng, X.; Zhang, W.; Kuipers, F.; Kema, I.; Barcaru, A.; Horvatovich, P. Dynamic binning peak detection and assessment of various lipidomics liquid chromatography-mass spectrometry pre-processing platforms. *Anal Chim Acta* **2021**, *1173*, 338674. DOI: 10.1016/j.aca.2021.338674.
(31) Krishnan, S.; Vogels, J. T. W. E.; Coulier, L.; Bas, R. C.; Hendriks, M. W. B.; Hankemeier, T.; Thissen, U. Instrument and process independent binning and baseline correction methods for liquid chromatography–high resolution-mass spectrometry deconvolution. *Analytica Chimica Acta* **2012**, *740*, 12-19. DOI: https://doi.org/10.1016/j.aca.2012.06.014.

(32) Urban, J. Resolution, Precision, and Entropy as Binning Problem in Mass Spectrometry. In *Bioinformatics and Biomedical Engineering*, Cham, 2018//, 2018; Rojas, I., Ortuño, F., Eds.; Springer International Publishing: pp 118-128.
(33) Pan, N.; Rao, W.; Kothapalli, N. R.; Liu, R.; Burgett, A. W.; Yang, Z. The single-probe: a miniaturized multifunctional device for single cell mass spectrometry analysis. *Anal Chem* **2014**, *86* (19), 9376-9380. DOI: 10.1021/ac5029038.
(34) Chen, X.; Sun, M.; Yang, Z. Single cell mass spectrometry analysis of drug-resistant cancer cells: Metabolomics studies of synergetic effect of combinational treatment. *Anal Chim Acta* **2022**, *1201*, 339621. DOI: 10.1016/j.aca.2022.339621.
(35) Holman, J. D.; Tabb, D. L.; Mallick, P. Employing ProteoWizard to Convert Raw Mass Spectrometry Data. *Curr Protoc Bioinformatics* **2014**, *46*, 13.24.11-13.24.19. DOI: 10.1002/0471250953.bi1324s46.
(36) Brenton, A. G.; Godfrey, A. R. Accurate Mass Measurement: Terminology and Treatment of Data. *Journal of the American Society for Mass Spectrometry* **2010**, *21* (11), 1821-1835. DOI: https://doi.org/10.1016/j.jasms.2010.06.006.
(37) Wang, Y.; Gu, M. The Concept of Spectral Accuracy for MS. *Analytical Chemistry* **2010**, *82* (17), 7055-7062. DOI: 10.1021/ac100888b.
(38) McCann, A.; Rappe, S.; La Rocca, R.; Tiquet, M.; Quinton, L.; Eppe, G.; Far, J.; De Pauw, E.; Kune, C. Mass shift in mass spectrometry imaging: comprehensive analysis and practical corrective workflow. *Analytical and Bioanalytical Chemistry* **2021**, *413* (10), 2831-2844. DOI: 10.1007/s00216-021-03174-1.
(39) Nguyen, T. D.; Lan, Y.; Kane, S. S.; Haffner, J. J.; Liu, R.; McCall, L. I.; Yang, Z. Single-Cell Mass Spectrometry Enables Insight into Heterogeneity in Infectious Disease. *Anal Chem* **2022**, *94* (30), 10567-10572. DOI: 10.1021/acs.analchem.2c02279.

Chapter 6 Advanced Mass Spectrometry Data Analysis Methods for Information Mining and Visualization

This is ongoing work collaborating with past and current group members from Dr. Zhibo Yang lab including Dr. Xiang Tian, Zongkai Peng, Yunpeng Lan, Tra Nguyen, Dan Chen, Amit Singh, Shakya Wije Munige, and Deepti Bhusal for data acquisition. The image fusion part, where I was involved, was included in a previous publication on multimodal imaging of amyloid plaques (Tian, X.; Xie, B.; Zou, Z.; Jiao, Y.; Lin, L.-E.; Chen, C.-L.; Hsu, C.-C.; Peng, J.; Yang, Z. Multimodal Imaging of Amyloid Plaques: Fusion of the Single-Probe Mass Spectrometry Image and Fluorescence Microscopy Image. *Analytical Chemistry* **2019**, *91* (20), 12882-12889.). The other parts were also included in a book chapter previously published with me as one of the co-1st authors (Tian, X.; Zou, Z.; Yang, Z. Extract Metabolomic Information from Mass Spectrometry Images Using Advanced Data Analysis. *Methods Mol Biol* **2022**, *2437*, 253-272.). In-house scripts were improved or prepared by me to handle SCMS or MSI data during collaboration.

6.1. Abstract

As MS is becoming one of the most powerful tools for bioanalytical analysis, more possibilities in mining MS data have been explored with new strategies from other fields of science to dig deeper into the insight of collected data. Connections have been made in order to convert the numbers from m/z values and intensities into data which shows clear biological or physical meanings. In this chapter, attempts with state-of-art data analysis methods applied on MS data are discussed in detail.[1]

6.2. Introduction

Mass spectra usually contains thousands of peaks as digitized signals, while the target for data analysis is the biological picture or physical meaning behind the numbers.[2] Therefore, finding useful information such as correlations among the observed spectral features or between spectral features and chemical substances is an important step for using mass spectrometry to answer biological questions. Traditional methods including PCA or Pearson Correlation are still widely adopted in recent studies, but there is an urge in the development and application of more advanced methods to better handle MS data with larger size and more complicated structure.[3, 4]

One of the major advantages of mass spectrometry against antibody-based methods would be its capability of untargeted analysis when hundreds of substances can be

extracted, ionized and analyzed at the same time.[5] However, the richness of information also leads to the dramatic increase in the number of dimensions included in the MS data and in the difficulty in finding information or pattern of interest. Therefore, dimensionality reduction is a commonly adopted strategy for MS data mining.[2] One of the most well-known techniques is Principal Component Analysis (PCA), which projects the original observations into the linear combination of a set of orthogonal principal components ordered by decreasing contribution towards the variance among original data.[3] Although PCA has advantages such as simplicity of implementation and can be run either with or without supervision, the applicability is limited by its linear nature.[6] To enhance the performance of dimensionality reduction, non-linear methods have been established and come into use, including t-distributed Stochastic Neighbor Embedding (t-SNE) in 2008[7] and Uniform Manifold Approximation and Projection (UMAP) in 2018[8]. The reduced data is more effective for clustering and classification by zooming into specific features.

Clustering and classification are two important tasks for mass spectrometry methods with spatial resolution. For MSI and SCMS experiments, similarity and dissimilarity among the signal from pixels or cells can lead to the discovery of important patterns and biological understandings. The simplest classification could be a fixed threshold value for binary classification, while for certain purposes a classification could be highly complicated with machine learning algorithms that

evolve with the given training set and can hardly be straightforwardly described with one single criterion composed with definite mathematical equations.[9, 10] Clustering and classification are commonly discussed at the same time because their common goal of finding certain groups among the whole set of data. The difference is that classification process is generally supervised, with the mandatory correct labels provided in the training set for the model to match and learn, while the clustering process do not have reference labels for the algorithm to learn from, thus setting it focus at finding the intrinsic difference between individual data points. There can also be a compromise between the two categories, with the strategy known as 'semi-supervised learning' to tune the learning accuracy.[11] Each algorithm has its own strengths and weaknesses with no universal solution for all sets of data, all depending on the specific need of the user and shape of data. Cluster Large Applications (CLARA)[12], Density-Based Spatial Clustering of Applications with Noise (DBSCAN)[13], and K-means[14] are typical unsupervised clustering algorithms, while Random Forest (RF)[15], Support Vector Machine (SVM)[16], and Neural Network (NN)[17] are commonly used supervised methods.

As a powerful analytical tool to study biological samples, mass spectrometry is commonly accompanied with other techniques providing information from other aspects, such as H&E staining or fluorescence microscopy.[18, 19] While different methods could have their advantages and specific targets, combination of information can inspire more understanding on the correlation between different

substances. When it comes to mass spectrometry imaging, the spatial resolution is generally limited by the size of each individual pixel, depending on the sampling technique and is typically at micrometer level. Whereas for microscopy images, the resolution is limited to the wavelength of the light, thus having much higher resolution but without chemical information. Image fusion of the two results can be performed to enhance the resolution of ion images or even to extrapolate the ion images and to predict the chemical profile of unscanned area.[20]

6.3. Methods

Raw data collected from mass spectrometers were converted into mzML and imzML format with ProteoWizard and imzMLConverter. Converted mzML and imzML files were imported into Python with pymzml and pyimzml packages. Data was read in forms of list, Numpy arrays, and Pandas data frames for storage, processing, and exportation. Signal was handled with Numpy and Scipy operations prior to dimensionality reduction, clustering and classification, and image fusion. For certain methods, data was performed in other programming languages including Matlab and R instead of Python. For dimensionality reduction, PCA was run with sklearn.decomposition.PCA package in Python, t-SNE was run with t-SNE package in R, UMAP was run with umap package in Python. Image fusion was performed with the standalone application.[20]

6.4. Results and discussion

The performance of advanced data analysis methods can be significantly influenced by the quality of data preprocessing steps, especially noise filtering and peak alignment. First of all, when the pre-processed data is oversized with unfiltered noise or misaligned peaks, the size of the data becomes much larger than expected, thus dramatically the cost for advanced methods is dramatically increased, even finally leading to the failure of the algorithm. Secondly, noise and misaligned peaks could severely impact the conclusion by bringing outliers of the model, or even draw biased conclusions when the amount of outlier is more than enough to create a change in the pattern of observed features. The image fusion performed was able to improve the resolution of the generated ion images towards a higher limit via comparison with the microscopy image.

6.5. Conclusion

Without proper annotations or validation with other methods, information gained from mass spectrometry data cannot directly provide answers to important and meaning biological questions. Advanced data analysis methods have significantly improved the efficiency of data mining in the massive amount of information from millions of MS peaks. With the combination of multiple tools, working pipelines

consists of data pretreatment, advanced data analysis, and database search can smoothly connect the patterns found in the MS data with related biological information.

6.6. Reference

(1) Tian, X.; Zou, Z.; Yang, Z. Extract Metabolomic Information from Mass Spectrometry Images Using Advanced Data Analysis. *Methods Mol Biol* **2022**, *2437*, 253-272. DOI: 10.1007/978-1-0716-2030-4_18.
(2) Chen, C.; Hou, J.; Tanner, J. J.; Cheng, J. Bioinformatics Methods for Mass Spectrometry-Based Proteomics Data Analysis. *Int J Mol Sci* **2020**, *21* (8). DOI: 10.3390/ijms21082873.
(3) Pearson, K. LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **1901**, *2* (11), 559-572. DOI: 10.1080/14786440109462720.
(4) Liu, H.; Chen, C.; Li, Y.; Duan, Z.; Li, Y. Chapter 9 - Characteristic and correlation analysis of metro loads. In *Smart Metro Station Systems*, Liu, H., Chen, C., Li, Y., Duan, Z., Li, Y. Eds.; Elsevier, 2022; pp 237-267.
(5) Zhang, X. W.; Li, Q. H.; Xu, Z. D.; Dou, J. J. Mass spectrometry-based metabolomics in health and medical science: a systematic review. *RSC Adv* **2020**, *10* (6), 3092-3104. DOI: 10.1039/c9ra08985c.
(6) Maaten, L. v. d.; Postma, E. O.; Herik, J. v. d. Dimensionality Reduction: A Comparative Review. 2008.
(7) Maaten., L. v. d.; Hinton., G. Visualizing Data using t-SNE. *Journal of Machine Learning Research* **2008**, *9* (86), 2579-2605.
(8) Leland McInnes, J. H., James Melville,. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv* **2020**, (1802.03426). DOI: https://doi.org/10.48550/arXiv.1802.03426.
(9) Meyer, J. G. Deep learning neural network tools for proteomics. *Cell Rep Methods* **2021**, *1* (2), 100003. DOI: 10.1016/j.crmeth.2021.100003.
(10) Zeng, W.-F.; Zhou, X.-X.; Willems, S.; Ammar, C.; Wahle, M.; Bludau, I.; Voytik, E.; Strauss, M. T.; Mann, M. AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics. *Nature Communications* **2022**, *13* (1), 7238. DOI: 10.1038/s41467-022-34904-3.
(11) González, C.; Astudillo, C. A.; López-Cortés, X. A.; Maldonado, S. Semi-supervised learning for MALDI–TOF mass spectrometry data classification: an application in the salmon industry. *Neural Computing and Applications* **2023**, *35* (13), 9381-9391. DOI: 10.1007/s00521-023-08333-2.
(12) Kaufman, L.; Rousseeuw, P. J. Clustering Large Applications (Program CLARA). In *Finding Groups in Data*, 1990; pp 126-163.
(13) Ester, M.; Kriegel, H.-P.; Sander, J.; Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon; 1996.

(14) Ikotun, A. M.; Ezugwu, A. E.; Abualigah, L.; Abuhaija, B.; Heming, J. K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data. *Information Sciences* **2023**, *622*, 178-210. DOI: https://doi.org/10.1016/j.ins.2022.11.139.
(15) Lebanov, L.; Tedone, L.; Ghiasvand, A.; Paull, B. Random Forests machine learning applied to gas chromatography – Mass spectrometry derived average mass spectrum data sets for classification and characterisation of essential oils. *Talanta* **2020**, *208*, 120471. DOI: https://doi.org/10.1016/j.talanta.2019.120471.
(16) Webb-Robertson, B. J. Support vector machines for improved peptide identification from tandem mass spectrometry database search. *Methods Mol Biol* **2009**, *492*, 453-460. DOI: 10.1007/978-1-59745-493-3_28.
(17) Seddiki, K.; Saudemont, P.; Precioso, F.; Ogrinc, N.; Wisztorski, M.; Salzet, M.; Fournier, I.; Droit, A. Cumulative learning enables convolutional neural network representations for small mass spectrometry data classification. *Nature Communications* **2020**, *11* (1), 5595. DOI: 10.1038/s41467-020-19354-z.
(18) Larson, K.; Ho, H. H.; Anumolu, P. L.; Chen, M. T. Hematoxylin and Eosin Tissue Stain in Mohs Micrographic Surgery: A Review. *Dermatologic Surgery* **2011**, *37* (8).
(19) Tian, X.; Xie, B.; Zou, Z.; Jiao, Y.; Lin, L.-E.; Chen, C.-L.; Hsu, C.-C.; Peng, J.; Yang, Z. Multimodal Imaging of Amyloid Plaques: Fusion of the Single-Probe Mass Spectrometry Image and Fluorescence Microscopy Image. *Analytical Chemistry* **2019**, *91* (20), 12882-12889. DOI: 10.1021/acs.analchem.9b02792.
(20) 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. *Nature Methods* **2015**, *12* (4), 366-372. DOI: 10.1038/nmeth.3296.