Spatially Resolved Proteomics of a Human Brain Tumour



Simon Davis

St Anne's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Spatially Resolved Proteomics of a Human Brain Tumour

Simon Davis, St. Anne's College
D.Phil Thesis, June 2021

Biological tissues contain diverse cell subpopulations crucial to their function, and proteins are the macromolecules that carry out many of these functions. Robust mass spectrometry-based methods that can accurately identify and quantify thousands of proteins within a sample have transformed the investigation of protein abundance and function within tissues and cells. However, mass spectrometry-based proteomics of tissue is traditionally performed on bulk samples. Therefore, information on the spatial composition and protein abundance patterns of those tissues is lost. Currently, the spatial context of proteins within a tissue is investigated using antibody-based high-resolution techniques with little multiplexing capacity.

There is a clear need for techniques that generate unbiased, quantitative, spatial expression profiles of many hundreds to thousands of proteins within a tissue. Presented within this thesis is the development of sample preparation, acquisition, and analysis workflows for spatially resolved proteomics using bottom-up mass spectrometry. This method was used to analyse an atypical teratoid/rhabdoid tumour (ATRT), a rare paediatric brain tumour.

Firstly, an optimised workflow was developed and used to characterise the proteomes of two neurons: Betz and Purkinje cells. Both cell types show highly specific proteomes. Purkinje cells show an increased abundance of several known Purkinje cell marker proteins. In addition, several novel candidate marker proteins for Betz cells were identified.
Secondly, the method was extended to allow for unbiased sampling of tissue in a spatially resolved manner. The method shows clear signals from a starting material of approximately 50 cells isolated from ATRT tissue. Analysis of the data within the tissue's spatial context highlights proteins and pathways showing spatially variable abundance within the tumour and adjacent normal tissue. This is performed without prior knowledge of the tissue composition by using spatially aware statistical methods.
Finally, orthogonal validation was performed using immunohistochemistry and mass spectrometry imaging which recapitulates the spatial variability observed in the bottom-up mass spectrometry approach.

Overall, the work presented here informs on methods for spatially resolved proteomics of tissue, laying the groundwork for further spatially resolved tissue proteomics and highlighting the utility of such approaches to understanding tissue biology.

Acknowledgements

First, I would like to thank my supervisors, Professors Roman Fischer, Olaf Ansorge and Benedikt Kessler, for providing the opportunity and resources to perform this research under their supervision. I am incredibly grateful for your guidance, encouragement, and enthusiasm.

All members of the Kessler group, past and present, have been an absolute pleasure to work with and enormously helpful. In particular, Darragh O'Brien, Hannah Jones, Adam Thorne, Adan Pinto Fernandez, Georgina Berridge, Raphael Heilig, Iolanda Vendrell, Svenja Hester, Zhu Liang and Andreas Damianou. You have all helped in making a fantastic, collaborative, and supportive environment.

Special thanks go to Connor Scott for his indispensable assistance with tissue sectioning and immunohistochemistry, and Phil Charles, for his expert guidance in bioinformatics. Thank you for always being willing to lend me your time. Thanks also to Janina Oetjen for the MALDI imaging data acquisition and guidance on its analysis.

Thanks also go to my friends in and outside of Oxford: Emma, Jake, Fran, Jenny, Chris, Felix, Peter, Ben, Dan, Andy, Leanne, Dan, Will, Phil and Kaine. Thank you for your friendship, advice and for all the good times.

I would also like to thank my family, especially my parents, who have always provided their unconditional encouragement, support and understanding over my journey, for which I am deeply thankful. And finally, I would like to thank Tasha for her love, patience, and everything.

Table of Contents

Spatially Resolved Proteomics of a Human Brain Tumour ii
Acknowledgements iii
Table of Contents iv
List of Figures and Tables viii
Abbreviations x
1 Introduction 1
1.1 Mass spectrometry-based proteomics 1
1.1.1 Bottom-up proteomics 1
1.1.2 Sample preparation 2
1.1.3 Liquid chromatography 5
1.1.4 Mass spectrometry 5
1.1.5 Identification of peptides and protein inference 8
1.1.6 Peptide and protein quantification 10
1.2 Laser capture microdissection 12
1.3 Spatial investigation of protein expression by mass spectrometry 16
1.3.1 Mass Spectrometry Imaging 17
1.3.1.1 Label-free mass spectrometry imaging 17
1.3.1.2 Labelled mass spectrometry imaging 19
1.3.2 Direct extraction from tissue 20
1.3.3 Tissue (micro)dissection 21
1.4 Spatial Statistics 22
1.4.1 Spatial autocorrelation 23
1.5 Atypical teratoid/rhabdoid tumours 26
1.5.1 Clinical characteristics of ATRT 27
1.5.2 Treatment of ATRT 28
1.5.3 Molecular characteristics of ATRT 28
1.5.3.1 SWI/SNF complex in ATRT 28
1.5.3.2 Molecular subtypes 29
1.5.3.3 Potential future therapies 30
1.6 Thesis aims 30
2 Development of a sensitive, quantitative method for LCM proteomics 31
2.1 Introduction 31
2.2 Materials and Methods 33
2.2.1 Tissue retrieval and processing 33

2.2.2 Laser-capture microdissection 34
2.2.3 Proteomic sample processing 34
2.2.3.1 Tissue lysis 34
2.2.3.2 In-cap digestion 34
2.2.3.3 In-solution digestion 34
2.2.3.4 Single-Pot Solid-Phase-enhanced Sample Preparation (SP3) 35
2.2.4 LC-MS/MS 36
2.2.5 Data analysis 36
2.2.6 Immunohistochemistry 37
2.3 Results and Discussion 39
2.3.1 Defining a simple starting point for LCM proteomics 39
2.3.2 Towards an optimised method for LCM proteomics 43
2.3.3 LCM proteomics reveals differences between human neuronal proteomes 52
2.3.4 Comparison to other methods 56
2.4 Conclusions 57
3 Spatially resolved proteomics of a human brain tumour 58
3.1 Introduction 58
3.2 Materials and Methods 59
3.2.1 Tissue retrieval and processing 59
3.2.2 Laser capture microdissection 60
3.2.3 Proteomic sample processing 60
3.2.3.1 Lysis, reduction and alkylation 60
3.2.3.2 SP3 60
3.2.3.3 S-Trap 61
3.2.4 LC-MS/MS 61
3.2.5 Mass spectrometry data analysis 63
3.2.5.1 Protein identification and label-free quantitation 63
3.2.6 Spatial data analysis 63
3.2.6.1 Feature-independent analysis 63
3.2.6.2 Feature-dependent analysis 64
3.2.7 Pathway analysis 65
3.2.8 Cell counting 66
3.3 Results and Discussion 66
3.3.1 Preliminary spatial proteomics sensitivity analysis 68
3.3.2 Pilot ATRT spatial proteomics experiment 70
3.3.2.1 Feature-dependent analysis 73

3.3.2.1.1 Proteomic maps 77
3.3.2.1.2 Feature-dependent pathway analysis 78
3.3.2.2 Feature-independent analysis 81
3.3.2.2.1 Missing values and dispersion-like patterns impact detection by the feature-independent method 85
3.3.2.3 High-throughput spatial proteomics 88
3.3.3 Improving the success rate of tissue capture 97
3.3.4 Spatial proteomic analysis of a replicate ATRT section 99
3.3.4.1 Feature-independent analysis 99
3.3.4.2 Feature-dependent analysis 102
3.3.5 Comparison between pilot and replicate experiments 103
3.3.5.1 Feature-dependent analysis 104
3.3.5.2 Feature-independent analysis 107
3.4 Conclusion 112
4 Further improvement, application and validation of spatially resolved proteomics 113
4.1 Introduction 113
4.2 Materials and Methods 113
4.2.1 LC-MS/MS 114
4.2.2 Data analysis 114
4.2.2.1 Protein identification and label-free quantitation 114
4.2.2.2 Experimental design and analysis of mass spectrometry parameter optimisation 115
4.2.2.3 Feature-independent analysis 115
4.2.2.4 Clustering 115
4.2.2.5 Pathway analysis 116
4.2.2.6 Cell-type map generation 116
4.2.3 MALDI Imaging 116
4.2.3.1 Trypsin digestion 116
4.2.3.2 MALDI matrix preparation 116
4.2.3.3 MALDI Imaging data acquisition 117
4.2.3.4 MALDI Imaging data analysis 117
4.2.4 Immunohistochemistry 118
4.2.5 Public data acquisition 118
4.3 Results and Discussion 119
4.3.1 Adaption of the existing protocol to 96 well plates 119
4.3.2 High throughput LC-MS/MS optimisation 120

4.3.2.1 Response surface optimisation of TIMS accumulation and ramp time 120
4.3.2.2 Response surface optimisation of intensity threshold and target 122
4.3.3 Short-gradient nLC/MSMS: A middle-ground for depth and throughput? 126
4.3.4 Spatial proteomics analysis of a large field of view 128
4.3.4.1 Feature-independent analysis 130
4.3.4.2 Clustering 134
4.3.4.3 Clustering-based pathway analysis 139
4.3.5 Posttranslational modifications 146
4.3.6 MALDI Imaging of ATRT tissue 147
4.3.6.1 Peptides 147
4.3.6.2 Lipids 151
4.3.7 Validation of spatial proteomic hits by immunohistochemistry 154
4.3.8 Reanalysis of public data 156
4.4 Conclusion 159
5 Summary and Future Perspectives 160
6 Appendix 164
7 References 193

List of Figures and Tables

Figure 1 – Overview of bottom-up proteomics workflow. 2
Figure 2 – Tandem mass spectrometry 8
Figure 3 – Peptide identification 10
Figure 4 – Label-free quantitation and downstream analysis 12
Figure 5 – Spatial autocorrelation testing using Moran's *I* 26
Figure 6 – Location of Purkinje cells within the cerebellum. 39
Figure 7 – Peptide and protein identifications from digest after recovering cells from the LCM cap. 40
Figure 8 – In-cap digestion workflow. 41
Figure 9 – Total ion chromatogram showing In-cap digestion improvement 42
Figure 10 – Peptide and protein identifications from the In-cap workflow. 43
Figure 11 – Schematic overview of the LCM workflow. 44
Figure 12 – Comparison of protein identifications of methods used for protein retrieval and digestion. 45
Figure 13 – Identification overlap of proteins detected in RIPA and TFE. 45
Figure 14 – Protein molecular weight distributions 47
Figure 15 – Protein isoelectric point distributions 48
Figure 16 – Protein GRAVY score distributions 49
Figure 17 – Summed peptide intensity of RIPA and TFE samples 50
Figure 18 – Quantitative performance of the tested methods 51
Figure 19 – Proteomic analysis of Purkinje and Betz cells 52
Figure 20 – Immunohistochemistry validation of proteins identified as differentially expressed between Purkinje and Betz cells 54
Figure 21 – Differentially expressed proteins correlated to public data. 55
Figure 22 – Spatially resolved proteomic workflow 66
Figure 23 – Protein identifications in ATRT processed with SP3 and S-trap 68
Figure 24 – Protein identifications on two LC-MS/MS systems from a range of ATRT areas. 70
Figure 25 – ATRT Histology and region used for raster sampling 71
Figure 26 – Distribution of proteins quantified per samples in the ATRT pilot experiment. 73
Figure 27 – Feature-dependent analysis of ATRT spatial proteomic data 75
Figure 28 – Spatial proteomic maps of four example proteins 78
Figure 29 – Functional analysis of feature-dependent ATRT proteomic data 80
Figure 30 – Overlap of significant proteins between feature-dependent and feature-independent approaches 83
Figure 31 - Example of feature-dependent significant proteins ‘missed’ by the feature-independent method. 87
Figure 32 – Feature-dependent analysis of high-throughput ATRT spatial proteomic data 90
Figure 33 – Schematic of ratio compression caused by imputing missing values 93
Figure 34 – High-throughput spatial proteomic maps of four example proteins 95
Figure 35 – Protein identifications from varying subpixel sizes 98
Figure 36 – Spatial proteomic maps of example proteins from a replicate tissue section 100
Figure 37 – Local Moran’s *I* spatial autocorrelation analysis of Histone H4. 102
Figure 38 – Feature-dependent analysis of a replicate tissue section 103

Figure 39 – Spatial proteomic maps of example proteins with different behaviour between experiments 106
Figure 40 – Overlap of spatially significantly differentially expressed proteins using the feature-independent method 107
Figure 41 – Density plots of quantified pixels per protein map 108
Figure 42 - Spatial proteomic maps of example proteins significant by feature-independent analysis in experiment one only 110
Figure 43 – Protein identified from a titration of ATRT area after high-throughput sample collection and preparation modifications 120
Figure 44 – Optimisation of instrument ion mobility parameters 122
Figure 45 – Optimisation of instrument PASEF MS/MS fragmentation intensity threshold and target 124
Figure 46 – Effect of PASEF MS/MS fragmentation targets on PSM quality 125
Figure 47 – Protein identifications on three LC-MS/MS systems from a titration of ATRT areas. 127
Figure 48 – ATRT slide for large field of view analysis 129
Figure 49 – Distribution of identified and quantified proteins in large region of ATRT 129
Figure 50 – Summed and mean intensities of each sample 131
Figure 51 – Spatial proteomic maps of example proteins in a large field of view 132
Figure 52 – Local Moran's *I* spatial autocorrelation analysis of haemoglobin and histone H4 133
Figure 53 - Spatial clustering of samples 135
Figure 54 – Spatial clustering of samples using Dynamic Tree Cut 137
Figure 55 - UMAP embedding of data 138
Figure 56 – Functional analysis of Clusters 1 and 3 140
Figure 57 - Functional analysis of Clusters 1 and 2 142
Figure 58 - Functional analysis of Clusters 6 and 13 144
Figure 59 – Spatial proteomic maps of immune cell marker proteins identified from cluster analysis 145
Figure 60 – MALDI imaging of ATRT peptides 150
Figure 61 – Candidate proteins for 1,375.75 m/z ion observed in MALDI imaging 151
Figure 62 – MALDI Imaging of ATRT lipids 153
Figure 63 – Immunohistochemistry validation of spatial proteomic maps 155
Figure 64 – Spatial proteomic analysis of public mouse uterus data. 158

Table 1 – Protein preparation steps for combinations of lysis buffers used in SP3 and In-solution digestion 38
Table 2 – Human Proteome Atlas immunohistochemistry data 56
Table 3 – Summary of four example proteins between LC-MS/MS systems and statistical methods 96
Table 4 – Proteins identified from ATRT tissue titration across LC-MS/MS system configurations 128
Table 5 – Peptides quantified in whole-slide raster with masses corresponding to 1,375 m/z peak in peptide MALDI imaging data 149
Table 6 – Summary of four example proteins from Piehowski et al. mouse uterus data 159

Abbreviations

ACN - Acetonitrile
AGC – Automatic gain control
ATP – Adenosine triphosphate
ATRT – Atypical teratoid/rhabdoid tumour
ASPH – Aspartate beta-hydroxylase
BMP – Bone morphogenetic protein
CID – Collision-induced dissociation
CHCA – alpha-cyano-4-hydroxycinnamic acid
CNS – Central nervous system
CSR – Complete spatial randomness
DAB - 3,3'-diaminobenzidine
DDA – Data-dependent analysis
DESI – Desorption electrospray ionisation
DF – Degrees of freedom
DHB – Dihydroxybenzoic acid
DIA – Data-Independent analysis
DMSO – Dimethyl sulfoxide
DTT – Dithiothreitol
ECM – Extracellular matrix
ESI – Electrospray ionisation
FACS – Fluorescence-activated cell sorting
FASP – Filter-aided sample preparation
FDR – False discovery rate
FFPE – Formalin-fixed paraffin-embedded
GASP – Gel-aided sample preparation
GRAVY – Grand average of hydropathy
GO – Gene ontology
GTP – Guanosine triphosphate
H&E – Haematoxylin and Eosin
HCD – Higher-energy collisional dissociation
HDAC – Histone deacetylase
HPA – Human protein atlas
IF – Intermediate filament
IHC - Immunohistochemistry
IMC – Imaging mass cytometry
ISD – In-solution digestion
LC – Liquid chromatography
LCM – Laser-capture microdissection
LESA – Liquid extraction surface analysis
LFQ – Label-free quantification
LE – Luminal epithelium
LPC – Laser pressure catapulting
nLC – Nanoflow liquid chromatography

MALDI – Matrix-assisted laser desorption/ionisation
MIBI – Multiplexed ion beam imaging
MRT – Malignant rhabdoid tumour
MS – Mass spectrometry
MS/MS – Tandem mass spectrometry
MSI – Mass spectrometry imaging
OCT – Optimal cutting temperature
OD – Optical density
PAGE – Polyacrylamide gel electrophoresis
PASEF – Parallel accumulation, serial fragmentation
PC – Phosphatidylcholine
PCA – Principal component analysis
PCR – Polymerase chain reaction
PEG - Polyethylene glycol
PRC2 – Polycomb repressor complex 2
PRPH – Peripherin
PTM – Post-translational modification
PYGB/L/M – Glycogen phosphorylase, brain/liver/muscle
Q-TOF – Quadrupole time-of-flight
RIPA - Radioimmunoprecipitation assay
RMS – Root-mean-square
RT – Room temperature
SDS - Sodium dodecyl sulphate
SHH – Sonic hedgehog
SIMS – Secondary ion mass spectrometry
SP3 – Single-pot solid-phase sample preparation
SS – Sum of squares
TBS/T – Tris-buffered saline with Tween
TEAB – Triethylammonium bicarbonate
TFA – Trifluoroacetic acid
TFE – Trifluoroethanol
TGFBI – Transforming growth factor-beta-induced protein ig-h3
TIMS – Trapped ion mobility spectrometry
TOF – Time-of-flight
UMAP – Uniform manifold projection
UV – Ultraviolet

1 Introduction

1.1 Mass spectrometry-based proteomics

Proteomics encompasses the study of the set of expressed proteins within the hierarchy of organelles through to whole organisms. Avenues for study can include levels of protein expression, localisation of protein expression, protein-protein interactions, post-translational modification of proteins and protein turnover/degradation[1]. Classical investigation into protein behaviour and function was performed with low throughput, antibody-based readouts such as western blotting, immunofluorescence, and immunohistochemistry, along with other biochemical and biophysical methods. Mass spectrometry initially showed promise for the study of proteins with the development of 'soft' ionisation techniques, which renders the protein molecules as intact ions in the gas phase, notably the development of matrix-assisted laser desorption ionisation and electrospray ionisation (MALDI & ESI)[2,3]. Today, these two techniques are now routinely used to characterise aspects of the proteome across many biological fields, along with other biomolecules such as lipids and metabolites. In the case of ESI, the advances in instrumentation, sample preparation, and data analysis techniques now allow for the comprehensive analysis of simple proteomes such as from yeast and for analysis depth of complex proteomes comparable to that achieved by RNAseq[4–8].

1.1.1 Bottom-up proteomics

The initial study of intact proteins by mass spectrometry – the 'top-down' approach – was challenging, leading to the development of the 'bottom-up' approach where proteins are initially enzymatically digested down to shorter peptides before separation by liquid chromatography coupled to a mass spectrometer with an ESI source (Figure 1)[9]. The presence, abundance and properties of the detected peptides are then used to infer information about the proteins they are derived from. This approach was helped dramatically by the large-scale DNA sequencing of organism genomes, allowing for the prediction of protein sequences from

genomes[10]. As various aspects of technology and method development occurred, the depth of coverage of the proteome has increased to a point where bottom-up proteomics can produce data with a depth comparable to modern sequencing-based transcriptomics methods.

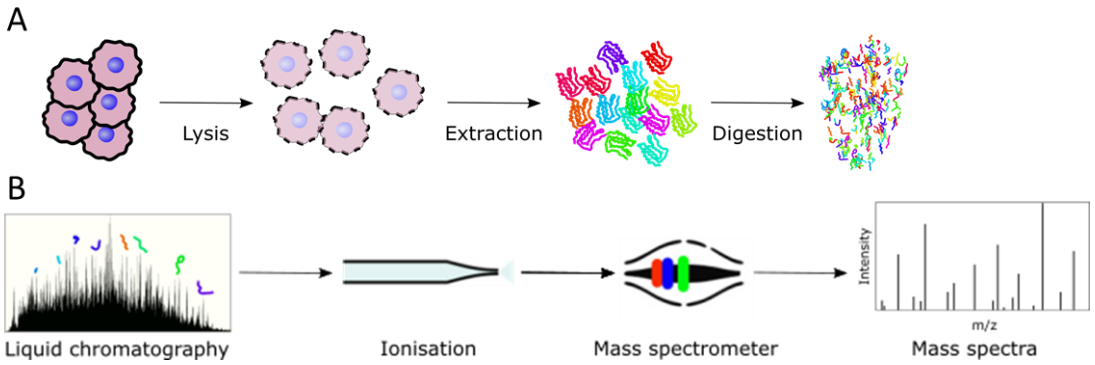

Figure 1 – Overview of bottom-up proteomics workflow.

(A) Cells and tissues are lysed and proteins extracted for digestion into peptides using a protease, commonly trypsin. (B) The resulting peptides are separated by liquid chromatography. Eluting peptides are ionised by electrospray ionisation and analysed by mass spectrometry.

1.1.2 Sample preparation

In the bottom-up workflow, the first step required is to release proteins from the material of interest through cell lysis. This step of cell lysis is an important one as the choice of lysis buffer will affect the set of proteins extracted into solution based on the properties of the detergents and/ chaotropes used[11]. For some sample types, the use of a mechanical disruption technique, such as sonication or bead-milling, is also required to disrupt material. Thought also needs to be given to other biomolecules captured during the extraction, such as genomic DNA, which can interfere with some sample preparation workflows and must be sheared or digested to fragments that do not interfere.

The next step generally involved the reduction of proteins' disulphide bonds using reducing agents such as dithiothreitol (DTT) or tris(2-carboxyethyl) phosphine (TCEP) to break disulphide bridges. The reduced cysteine thiol groups are then capped with an alkylating agent

such as iodoacetamide (IAA), which prevents the reformation of disulphide bridges after removing the reducing agent.

Trypsin is the most used enzyme for proteolytic digestion of proteins to peptides before LC-MS analysis[12]. Trypsin has a high specificity for cleavage at the C-terminal side of arginine and lysine, generating peptides of masses and charge states that are amenable to analysis by mass spectrometry[13]. For some proteins without many arginine or lysine residues, trypsin is less useful because the peptides generated would be larger than the optimal size. In this case, other enzymes can enhance the sequence coverage. Examples of these complementary enzymes are elastase, chymotrypsin, and Glu-C; these can be used alone or in combination with trypsin and each other to boost sequence coverage and depth[4,14]. Histone proteins are an example where trypsin will generate many peptides that are too short, and approaches such as using propionic anhydride derivatisation of lysine residues have been developed to improve sequence coverage of the lysine-rich tails[15].

Depending on the protein extraction method, a sample clean-up step may be required before protein digestion to remove chemicals that are detrimental to the enzyme's activity. A critical early method for MS-based bottom-up proteomics employed 1D SDS-PAGE (sodium dodecyl sulphate-polyacrylamide gel electrophoresis), allowing for clean-up and reactions to be performed on the proteins trapped within the gel[16]. This method also has the benefit of fractionating the proteins, reducing the complexity of each sample as the gel is cut into smaller pieces that are analysed separately. A disadvantage of the in-gel approach is its low throughput. In-solution digestion of proteins has a much higher throughput but will often require clean-up of the extracted protein. Methods based on precipitation, filtration, trapping protein within polyacrylamide gels and aggregation/affinity have been developed and employed to remove detergents and chaotropes from samples before protein digestion. Protein precipitation either by methanol-chloroform or acetone is commonly used before in-

solution tryptic digestion[17,18]. Filter-aided sample preparation uses molecular weight cut-off filters to retain SDS denatured protein and remove SDS by multiple washes of urea – a chaotrope that trypsin is tolerant of in moderate concentrations (< 1 M)[19]. Gel-aided sample preparation is a modification of the in-gel approach where the acrylamide is polymerised in the presence of the protein sample and then shredded to allow for the processing in a single tube[20]. S-TRAP is a hybrid method that uses both protein precipitation and the filtration of those precipitates to allow high concentrations of SDS and its subsequent removal[21]. Single-pot solid phase sample preparation (SP3) and protein aggregation capture (PAC) are methods that rely on denatured protein aggregating around micron-scale particles (mainly magnetic beads) in solution to allow for detergent, chaotrope, contaminant removal and volume reduction prior to digestion[22,23]. As the protein remains adsorbed or aggregated to the surface of the beads throughout most of the procedure, there is less opportunity for sample loss.

The peptides generated by the proteolytic digest also often require clean-up before MS analysis to remove contaminants such as salts that can impair or interfere with the LC-MS system. Solid-phase extraction systems using C18 material are standard as this is the same solvent and sorbent system used in the chromatography before MS analysis[24]. The SP3 method also contains a peptide clean-up step where the generated peptides bind to the magnetic beads, allowing for contaminant removal by washing with acetonitrile.

The proteomics community lacks a single standard method for protein extraction and digestion, partly due to the wide range of input materials with different properties and requirements. In addition, some methods are more easily scaled to higher-throughput 96-well formats or automated approaches, so the time efficiency of a method will also become an essential factor as the number of samples processed increases. Often, existing methods are optimised by the user for their specific application – for example, volumes can be minimised in in-solution digestions when working with low-abundant samples to reduce sample loss to

plasticware – something that is generally not a concern if working with a more abundant sample.

1.1.3 Liquid chromatography

The following process in bottom-up mass spectrometry is the separation of peptides by liquid chromatography (LC) to reduce the complexity of the mixture entering the mass spectrometer. Typically, this is done using reverse-phase chromatography on columns packed with C18-coated beads at acidic pH[25]. Peptides are loaded onto the column in a highly aqueous solvent mix (e.g., 2 % acetonitrile, 98 % water, 0.1 % formic acid), resulting in peptides adsorbing to the C18 material through hydrophobic interactions. The proportion of acetonitrile is then increased as time passes, disrupting the peptide-C18 interactions and causing elution of the peptides over time based on the strength of their interaction with the C18 material. Shorter, more hydrophilic peptides will elute earlier in the gradient, while longer and more hydrophobic peptides will elute later in the gradient[26]. Typically, using longer chromatographic gradients will result in fewer peptides reaching the mass spectrometer at any given time, allowing more peptides to be detected[27]. As mass spectrometers have become capable of sequencing peptides at a higher speed, shorter gradients have become more valuable as there is less compromise on the depth of coverage[28–32]. Recently, a novel high-pressure LC system – the EvoSep One – was developed where samples are first loaded onto disposable C18 pre-columns, resulting in a more robust system as contaminants do not build up in the system over time [33]. Additionally, the gradient is pre-formed at low pressure before being pushed through the C18 column with a single, high-pressure pump – this also results in low overhead and increased robustness.

1.1.4 Mass spectrometry

Mass spectrometry relies on the interaction of ions with electromagnetic fields to manipulate and measure those ions' mass-to-charge ratios (m/z). In ESI, peptides are ionised into the gas

phase at a narrow tip where a voltage is applied[3,34]. This emitter tip can either be integrated with or connected to the end of the column. The generated ions are then drawn into the inlet of the mass spectrometer as it is held at a rough vacuum. The ions are then accelerated through the mass spectrometer by a series of multipole electrodes that can also act to filter ions based on their m/z through setting field strengths to only allow stable paths for a subset of ions. Then, ions can be mass analysed and detected in a wide range of analysers, such as time-of-flight (TOF) tubes and ion traps. TOF-based analysers measure the time taken for an ion to move through a field-free region before impinging upon a detector – ions are all accelerated with an equal force, so their travel time is proportional to their m/z ratio. The arrival time is used to generate a mass spectrum based upon calibration with known standards. Ion traps rely upon first trapping ions within a low potential and altering that potential to selectively eject ions from the trap, where they impinge upon a detector[35]. The Orbitrap is both an analyser and detector as ions enter and proceed to orbit around a central electrode and oscillate along its length at a frequency proportional to their m/z ratio. These oscillations induce a current within the outer electrodes, which is detected and then Fourier-transformed into the frequency domain and then converted to a mass spectrum[36].

Ions detected in a mass spectrum by the instrument can then be isolated and subjected to a range of fragmentation techniques that generate fragments of ions prior to mass analysis, allowing for structural information to be determined. The intact ion that is fragmented is termed the precursor ion. In the case of peptides, this is usually a form of collision-induced dissociation (CID) with a neutral carrier gas such as nitrogen or helium as they pass through a multipole or in an ion trap. In the case of peptide ions, collisions with the gas induce fragmentation along the peptide backbone at the peptide bonds between the amino acids[37]. These fragmentations will generally occur at most peptide bonds, generating a ladder of *b and y* ions (n-terminal and c-terminal charged fragment, respectively). The mass spectrometer then analyses these ions as before to generate fragment ion spectra or MS2 spectra (Figure

2)[38]. Instruments with ion traps can perform multiple rounds of isolation and fragmentation if desired. For example, when using isobaric tags, the reporter ions can be detected following a subsequent fragmentation of the *b* and *y* ions detected in an MS2 spectrum[39]. Other forms of fragmentation that rely on gas-phase chemical reactions (electron transfer dissociation (ETD)) or laser-based transfer of energy (UV photodissociation) generate ions other than *b* and *y*, which provides complementary information that can help with the identification of post-translational modifications[40–42].

Another property of ions that some MS instruments can measure is their collisional cross-sectional area (CCS), a property determined by an ion's shape and size in the gas phase[43]. In the case of trapped-ion mobility spectrometry (TIMS), this is done by holding ions back from proceeding further into the MS using an electric field while gas pushes them into the instrument[44]. The electric field is ramped down, releasing ions in descending order of CCS as larger ions receive a larger force from the gas and overcome the electric field first. With two of these ion mobility cells, ions are released by their mobility while the next packet of ions for analysis is accumulated, allowing for increased sensitivity[45]. The millisecond timescale of this ion-mobility technique meshes well with the microsecond timescale of TOF analysers, allowing for many TOF spectra to be acquired during an ion elution ramp[46]. As a peptide's ion mobility is correlated with its mass and charge state, and tryptic peptides have a charge state > 1 +, uninformative 1+ ions can be excluded from being chosen for isolation and fragmentation[47,48].

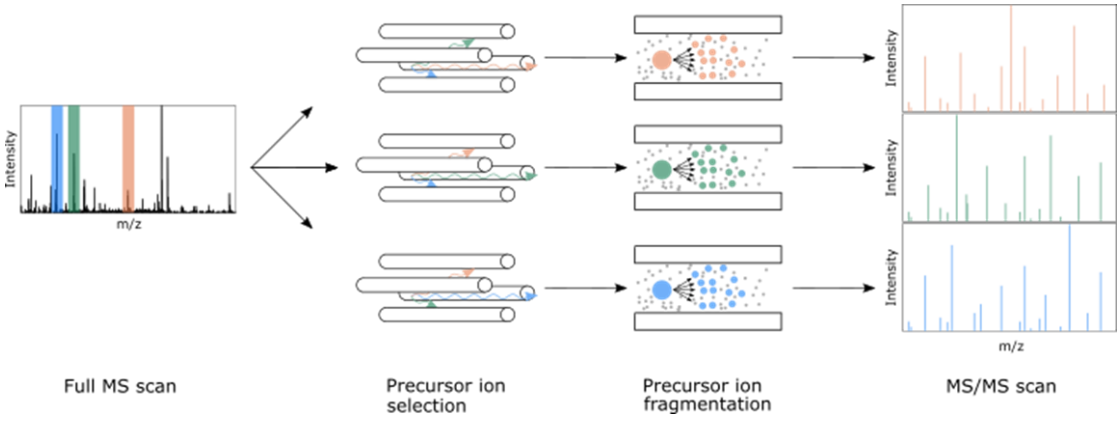

Figure 2 – Tandem mass spectrometry

Peptide precursor ions detected in a full MS scan (MS1) are fragmented to obtain sequence information. Precursor ions are selected for fragmentation by filtering ions outside of a narrow m/z range that contains the precursor. These precursor ions are then fragmented and mass analysed to generated fragment ion spectra.

1.1.5 Identification of peptides and protein inference

The mass spectra measured after peptide fragmentation and the mass of the peptide (calculated from the m/z of the precursor ion) are used to identify the peptides present within a sample. From this set of peptides, a non-redundant set of protein groups is determined following the parsimony principle where each protein group will have at least one sequence-unique peptide, specific to only that group in the context of the database used for comparison[49].

Many computational tools have been developed to analyse MS2 spectra generated after fragmentation of precursor ions to assign amino acid sequences to spectra. These tools fall into two categories: de-novo sequencing and database matching. The *de novo* sequencing approach attempts to assign amino acid m/z to the m/z differences between peaks in the spectra, leading to either an incomplete or complete peptide sequence[50]. *De novo* sequencing's advantage is that it does not require prior knowledge of gene/protein sequences and can be used to identify novel peptide sequences, such as monoclonal antibody characterisation[51]. Database matching through the use of 'search engines' is the more

commonly used approach (Figure 3)[52]. The m/z, charge state and fragmentation spectrum of each precursor ion are compared to a database containing theoretical spectra generated by an *in-silico* digest of protein sequences (usually the organism's entire set)[53,54]. The set of theoretical spectra is also dependent on the specified protease's specificity, the number of allowed missed cleavage sites per peptide and any specified post-translational modifications (PTMs). For each acquired fragment spectrum, the search engine will compare it to the set of theoretical spectra whose precursor mass would fall within a tolerance window that is adjusted based on the mass analyser's mass accuracy. The search engine generates a score for each comparison, representing the quality or likelihood of each match. These peptide-spectrum matches (PSMs) could, however, be a result of a false-positive. The most common approach to control the level of false assignments is to perform the database search process again, but this time against a set of 'decoy' sequences generated from structured rearrangement of the real or 'target' protein sequences – termed the 'target-decoy' approach[55]. As these decoy sequences should not be present within the sample, any matches to these sequences can be considered false positives and are used as proxies for false-positive target hits. The distributions of the target and decoy scores are then compared, and a score threshold set to limit the inclusion of decoy peptides to a specified false discovery rate (FDR), typically 1 %. To further differentiate between true-positives and false-positives, other feature distributions can be used to distinguish between them, such as peptide length, missed cleavages and charge states[56,57].

As bottom-up proteomics measures peptides, the set of identified peptides need to be assembled to infer the presence of proteins within the original sample[49]. An identified peptide will not necessarily be unique to one protein and might be shared between multiple proteins. These shared peptides do not allow for explicit identification of the proteins they are generated from. Protein inference methods attempt to infer the proteins most likely to be present within a sample based on the observed peptide sequences[58–61]. As some of these

protein identifications will be ambiguous, they can be collected to form protein groups. A protein group is a set of proteins that could be present within the sample, but they are indistinguishable from one another based on the set of peptides identified. Often these protein groups will contain isoforms of the same gene but can also contain proteins derived from different genes. These protein groups will be referred to as proteins for simplicity, but where a protein group's members are from different genes, these will be denoted using a semi-colon.

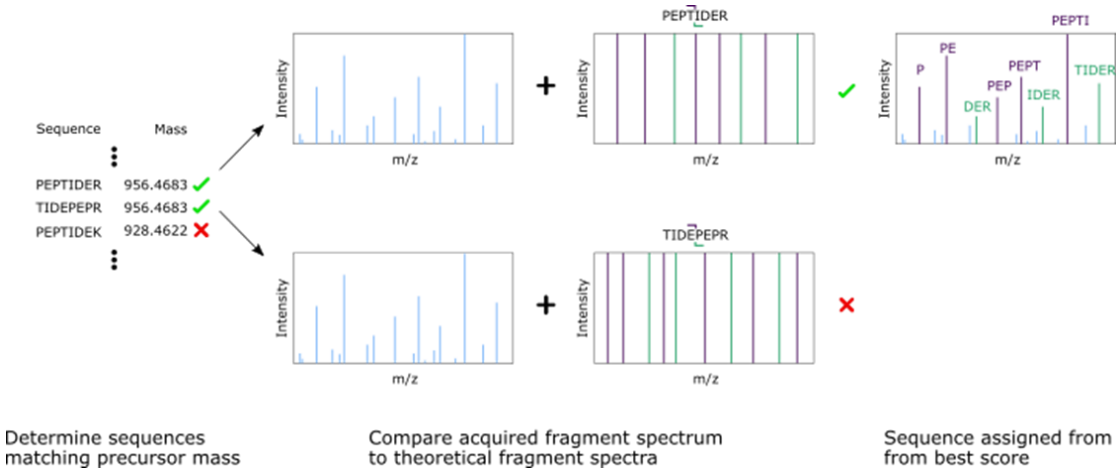

Figure 3 – Peptide identification

Precursor ion masses are compared to an *in-silico* digest of the analysed proteome. For each precursor, fragment ion spectra are compared to theoretical fragment spectra from sequences that match the precursor ion mass. The search engine assigns scores to each match, and the best scoring match is chosen as the peptide sequence identity.

1.1.6 Peptide and protein quantification

While the presence and absence of proteins within a sample are of interest, knowledge of their abundance within the sample is perhaps more interesting and valuable, as protein expression in response to stimuli is not necessarily a binary on/off event. Instead, these responses manifest as changes in protein concentration within cells, tissues and biofluids. Quantitative proteomics methods either try to determine protein copy numbers or concentrations within the same sample, termed absolute quantitation, or define relative ratios or abundances of each analyte across multiple samples, termed relative quantitation[9]. Mass

spectrometry is not an inherently quantitative method as the response of an analyte depends on its physical and chemical properties[62]. This property results in most quantitative proteomics experiments comparing the abundance of the same peptide across different samples, with absolute quantification methods relying on comparison to internal or external standards of known concentrations. Both approaches can be performed using metabolic or chemical labels to multiplex samples or using a label-free approach where each sample is analysed individually. The quantification strategy used in this thesis is (relative) label-free quantification (LFQ) (Figure 4). A range of LFQ analysis tools and algorithms have been developed, for example, MaxLFQ, IonQuant, FlashLFQ and Skyline[63–66]. In general, peptide intensity profiles from consecutive (or near-consecutive) MS1 spectra are assembled into 3-dimensional m/z, retention time, and intensity features, or 4-dimensional m/z, retention time, mobility, and intensity features (on instruments with an ion-mobility analyser)[52]. These features can then be aligned across multiple samples based on retention time (and ion-mobility if available) to handle retention-time shifts over the time taken to acquire data. These features' intensity is calculated and generally normalised between samples to account for sample processing and other sources of variability. As peptide identifications can be attached to these peptide features, and features are aligned to each other, the transfer of a peptide sequence identity from a run where sequence information was obtained to a run where the peptide was not sequenced but is still present in the MS1-feature data can be performed[67,68]. This identification transfer process allows the recovery of quantitative information when peptides are present but not sequenced (e.g. because of poor or no spectra of that peptide being acquired). In the MaxQuant software used in this thesis, this process of identification transfer is termed 'match-between-runs'.

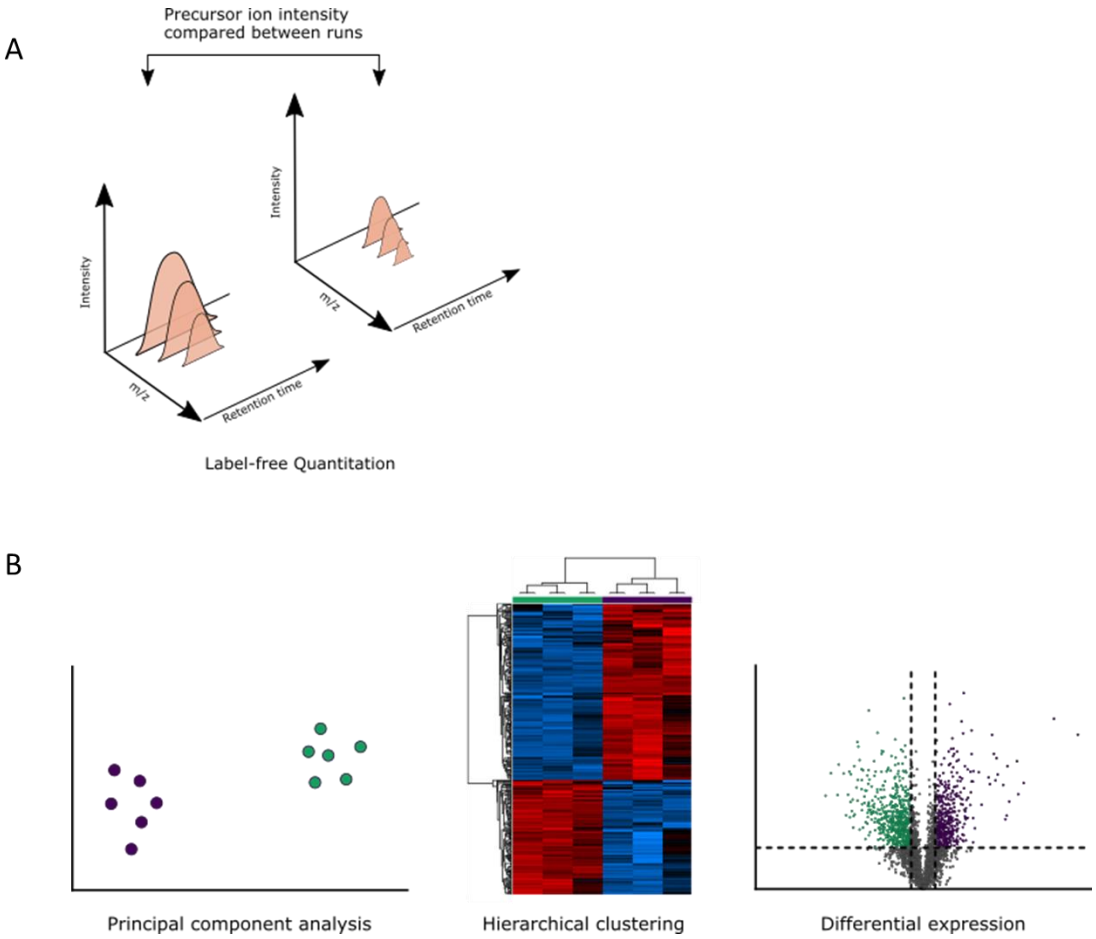

Figure 4 – Label-free quantitation and downstream analysis

(A) Peptides eluting over the retention time are collected into multidimensional features. These features are then quantified by comparing the intensity of the same feature across multiple samples. (B) After determining the intensity of proteins from their sets of peptides, these protein intensities can be used for downstream analysis. Typical analyses include dimensionality reduction, such as principal component analysis, hierarchical clustering and differential expression testing via *t*-tests and ANOVA.

1.2 Laser capture microdissection

Laser capture microdissection (LCM), also known as laser microdissection (LMD), is used to isolate subpopulations from tissue sections. LCM systems are based on light microscopes equipped with a laser to allow for the collection of specified regions of the tissue section through a range of capture devices.

The use of a laser for microdissection was first described for the microdissection of chromosomes before micro-cloning experiments[69]. The addition of a capture device was first described by Emmert-Buck et al. and consisted of applying a thin, transparent thermoplastic

film to the tissue slide[70]. A low energy infrared laser is used to melt the film in locations containing cells and tissue of interest. The melted film cools and bonds to the material of interest, and the process is repeated across the slide. Once complete, the film can be removed, bringing material of interest with it as that material adheres more strongly to the film than the glass slide. The film is then placed within a microcentrifuge tube containing a buffer to isolate the biomolecules of interest.

An ultraviolet light-based method was later developed, which relies on mounting the tissue section on a glass slide coated with a thin membrane[71]. These systems use the ultraviolet laser to ablate a perimeter of tissue surrounding the tissue of interest. Once the area of interest is separated from its surrounding, current systems increase the laser power to either catapult the tissue into a collection device placed above or below the tissue. An advantage of this type of system is the rapid, non-contact collection of samples. In an infrared, film-based system, tissue can non-specifically adhere to the film and cause sample contamination. The separation of the tissue of interest by ablation in an ultraviolet-based system reduces the chance of unwanted tissue contaminating a sample, a feature of great importance when isolating low numbers of cells. In addition, these systems can collect material faster and in a more high-throughput compatible fashion. These systems can also collect material from uncoated glass slides as the above process used to catapult tissue can also catapult tissue straight from the slide, but with much lower precision. This is the type of LCM system used in the work presented in this thesis.

LCM has been applied to a wide range of sample types, including frozen tissue, FFPE (formalin-fixed paraffin-embedded) tissue, plant tissue, live cell cultures, forensic samples, chromosome spreads, stained and unstained alike for the analysis of DNA, RNA, protein, lipids, carbohydrates, and small molecule metabolites[72–79]. The preparation of the tissue does present a dilemma. On the one hand, frozen tissue has excellent preservation and recovery of

biomolecules but has limited histological detail and morphology preservation. On the other hand, FFPE presents real challenges for preserving and extracting many biomolecules but is the gold standard for histological detail and preservation of tissue morphology due to extensive formalin crosslinking[80]. In the context of extraction of proteins for proteomics, a wide range of sample preparation methods have been developed for the analysis of FFPE tissue and are primarily benchmarked against equivalent extractions from frozen tissue, with recent techniques reaching a yield of approximately 95 % from FFPE tissue in comparison to frozen[81,82]. However, these were performed in bulk tissue and workflows that produce gains in bulk tissue do not necessarily translate to samples with low input amounts.

Due to the lack of a coverslip, optical resolution is limited and results in a loss of visible morphological detail even when using FFPE tissue. As the goal of LCM is to isolate specific cells and tissue structures of interest, these features do need to be visualised and identified by the operator, using the microscope either by light microscopy or by using fluorescent reporters. Some cellular and tissue features, such as precancerous cells, may require a trained pathologist for identification. Therefore, the coupling of automated image analysis to LCM systems could prove to be a powerful combination. Mund et al. [83] recently developed the concept of Deep Visual Proteomics by combining high-resolution imaging, machine-learning-based image analysis and a sensitive proteomics workflow. They first classify cells into subtypes using a trained machine learning model on images obtained using a high-resolution scanning microscope and annotate the edges of these cells. These annotations are then transferred onto the LCM system, and cells are automatically cut and captured based on their classification before proteomic analysis. Investigating the spatial heterogeneity of melanoma, they isolated cancer cells from the inner tumour and cancer cells on the periphery of the tumour and show that the central cells have a prognostically unfavourable proteome state compared to the peripheral cells.

LCM has been applied to study proteomes in a range of organisms, tissues, and disease states. Early studies utilising mass spectrometry were limited by a large amount of time required to collect enough sample. Banks et al. first demonstrated the potential of generating samples for proteomic analysis using LCM by showing the two-dimensional electrophoretic and mass spectrometry proteomic profiles (of 9 identified proteins) remained unchanged after exposure to the LCM laser, and that isolated cervical epithelial cells displayed a different two-dimensional gel electrophoresis profile to that of bulk tissue[84]. These findings were reinforced the following year when Emmert-Buck et al. performed a similar experiment and showed that proteins visible on a 2D gel from microdissected tissue were present at the same molecular weight and isoelectric point as in bulk tissue. Additionally, they found large differences in the 2D-PAGE profiles of epithelial and stromal cells[85].

Since its first use for protein investigation, LCM has investigated the proteome of a range of organisms, tissues, and disease states. Non-mammalian applications include plant vasculature and roots, parasite intestine and parenchyma, and chicken egg cuticle, all using LCM to isolate specific tissue of interest from surrounding tissue. In the context of human disease, the most extensive use of LCM-coupled proteomics has been in cancer research, predominantly for its utility in separating tumour cells from the surrounding tissue. De Marchi et al. [86] used LCM to identify and quantify more proteins from isolated breast tumour cells compared to a whole tissue lysate. Those proteins were more specific to the epithelial cells isolated, whereas the whole tissue lysate showed a higher level of stromal and blood proteins. This is exemplified by the LCM samples showing significant differential expression of the estrogen receptor 1 within a cohort of 38 cases with good and poor outcomes. This difference was not observed in the whole tissue lysate analysis. Großerueschkamp et al. [87] used LCM-proteomics in combination with Fourier transform infrared imaging to isolate tumour cells from two subtypes of diffuse malignant mesothelioma to a depth of over 3,000 proteins. Within the proteins differentially

expressed between the two subtypes, five proteins are known markers used clinically to distinguish between the subtypes.

1.3 Spatial investigation of protein expression by mass spectrometry

Tissues contain a diverse range of cell types and subpopulations. The location of cells within a tissue and their spatial neighbourhood is crucial for determining their identity and function[88–93]. The cellular composition of tissue has substantial effects on measured coexpression signals within the molecular profiles of bulk, and changes in cellular tissue composition may contribute to different disease outcomes[94]. Changes in the composition of the surrounding microenvironment could also contribute as cell functions are influenced by signals from the microenvironment[95]. In tumours, complex signalling occurs between tumour and normal cells that have been co-opted to promote tumour cell function and survival[96,97].

Tumour heterogeneity is an obstacle for cancer treatment and research into tumour biology[98–100]. Most cancer sub-typing is based on inter-tumour comparison of molecular profiles to identify common mutations and expression alterations[101–106]. However, this approach does not account for the heterogeneity of cell types and cellular states within each tumour. Tumours have been shown to have high degrees of genetic heterogeneity, and it would be expected that this genetic heterogeneity would result in heterogeneity of protein expression [107–112]. In addition to genomic factors, non-genomic factors such as the extracellular matrix, immune cell components, hypoxic state, and sub-clonal cooperation can drive progression and heterogeneity[107]. A new understanding of the interaction between the constituent cells of the tumour environment could aid the discovery of new biomarkers and therapeutic targets or patient-specific therapy usage.

Recent technology developments in DNA and RNA sequencing technologies now allow for the generation of near-complete genomes and transcriptomes. This has furthered the understanding of many aspects of biology from the level of cells through to populations and in

many human disease contexts[113–115]. Furthermore, the development of technologies capable of retaining the spatial context of these genomic and transcriptomic profiles have enabled the characterisation of spatial heterogeneity of these profiles within a tissue[116,117]. However, while genomic and transcriptomic alterations may act as drivers of disease, it is the proteins they code for that regulate essentially all cell processes[9]. A range of mass spectrometry-based techniques are available to map the distribution of proteins throughout tissues and cells.

1.3.1 Mass Spectrometry Imaging

Mass spectrometry imaging (MSI) enables the determination of the spatial distribution of molecules within a sample by rasterising an x, y grid over the sample with a focussed ionisation source. MSI can be used in both targeted and untargeted approaches. Untargeted approaches use ionisation techniques such as MALDI and desorption electrospray ionisation (DESI) to desorb, ionise and detect the molecules within a sample using a range of mass analysers. Targeted approaches such as imaging mass cytometry (IMC) and multiplexed ion beam imaging (MIBI) use metal isotope-tagged antibodies against the protein or RNA molecules of interest[118,119]. The metal isotopes are then liberated from the tissue, ionised and analysed to quantify the isotope abundance. In both MSI approaches, the molecule or reporter abundance can be mapped back to the original coordinates producing an image for each detected ion which can be registered with optical images of the same sample. Combining the MSI data with histology is vital for data interrogation and interpretation by placing the data within a known context of tissue structure and organisation[120].

1.3.1.1 Label-free mass spectrometry imaging

Initial MSI experiments were performed using secondary ion mass spectrometry (SIMS), where an ion beam is rastered across the sample surface, generating secondary ions from the sample, which are then introduced into the mass spectrometer[121,122]. With SIMS, there is often significant fragmentation of the molecules analysed, complicating the interpretation of the generated spectra[123]. The use of MALDI as a soft-ionisation technique in MSI allows for the

imaging of larger biomolecules such as intact proteins, as well as metabolites, lipids, and peptides generated by *in-situ* digestion of proteins[121,124]. In MALDI-based MSI, a matrix compound is applied to the tissue to transfer the laser energy to the analytes and ionise them. The spatial resolution of MALDI imaging is determined by the diameter of the laser and the size of the crystals formed by the matrix compound[125]. In DESI, the tissue is probed using a stream of charged solvent that causes analyte desorption and ionisation[126]. An advantage of DESI is that no upfront sample preparation is needed to prepare the tissue for imaging. SIMS can achieve the highest resolution of ≥ 20 nm of the three techniques, followed by MALDI at ≥ 600 nm (with current commercial systems achieving 5-10 µm) and DESI at ≥ 10 µm[122].

Most instruments used for MADLI-MSI are based on TOF mass analysers due to their high mass range, throughput, and mass resolution and accuracy. To confidently identify most analyte molecules detectable in MSI experiments, they must be fragmented to gain information about their structure. Performing MS/MS analysis for ions across the tissue adds a significant amount of analysis time to an already lengthy imaging process. Therefore, fragmentation is often not performed in the initial imaging experiments[127]. The molecular weight of the detected (intact) ions can be compared to molecule weight values in a database generated from chemical structures or orthogonal approaches where identification can be performed, such as LC-MS/MS. In the case of lipids, the measured molecular weight can be matched to public databases using a narrow mass tolerance window to obtain a putative identification of the class of lipid. However, to identify which member of the lipid class is present, fragment spectra need to be acquired, likely in subsequent experiments, and compared to that of a standard compound[128]. For peptides after *in-situ* digestion, the molecular weight of the detected peptides can be matched to the list of peptides identified in a bottom-up proteomics experiment[127]. Additionally, a subset of ions of interest could be targeted for fragmentation in follow-up MSI experiments and identified through matching parent ion mass and fragment ion masses in a database search as in bottom-up proteomics. However, identifying peptides from

tissue imaging is challenging due to the extreme complexity of the peptide mixture present in each pixel due to the lack of up-front separation.

1.3.1.2 Labelled mass spectrometry imaging

In an attempt to address the throughput of multiplexed immunofluorescence, two similar targeted MSI techniques have been developed that use antibodies labelled with heavy metal isotopes. These techniques, imaging mass cytometry (IMC) and multiplexed ion-beam imaging (MIBI), involve probing a tissue section with a cocktail of antibodies labelled with approximately 40 heavy metal isotope tags[118,119]. In IMC, tissue is ablated from the section by a laser in a raster and atomised, releasing the heavy metal isotope tags, which are then ionised by another laser. In MIBI, the ions analysed are secondary ions produced from firing an ion gun at the tissue. The metal ions are then introduced into a TOF mass spectrometer, detected and quantified. The abundance of the metal ions in each ablated pixel can be mapped back to the spatial coordinates and visualised to show the distribution of its antibody's binding across the tissue, generating multiplexed images similar to a multiplexed immunofluorescence image. Both methods have similar resolutions; IMC uses a laser with a 1 µm diameter, and MIBI's primary ion beam diameter is 500 nm. These high-resolution techniques generate spatial information at subcellular resolution and can investigate interdependencies between complex mixtures of cell populations within healthy and diseased tissue[129]. As both techniques are fundamentally antibody-based, they are compatible with FFPE tissue – assuming the individual antibodies perform well in FFPE tissue and the metal isotope conjugation does not alter the antibody performance. As there are vast amounts of FFPE tissue samples in pathology archives, techniques compatible with samples preserved by FFPE represent a valuable tool for clinical research[130,131].

These two techniques have mainly been applied in the context of cancer and immune profiling[129]. The initial work was performed in tumour tissue and showed good agreement with known pathology and literature findings. IMC and MIBI have been applied to the investigation

of tumour microenvironment signatures[132]. Keren et al. found two spatial phenotypes of immune cell infiltration: compartmentalised and mixed. Compartmentalised samples were predominantly comprised of regions of either tumour or immune cells, with the mixed phenotype shows a high degree of mixing of tumour and immune cells. The compartmentalised phenotype correlated with increased survival. Keren et al. also show that tumours with similar amounts of immune cells can differ in their spatial organisation. Spatial features such as this dichotomy highlight the benefit of obtaining spatially resolved data. Another noteworthy application of IMC was the analysis of a large breast cancer cohort[132]. Ali et al. found that different cell-cell interaction profiles are associated with established genomic subtypes[133]. These cell-cell interaction profiles are based on cell-cell proximities, a feature captured with this type of high-resolution spatial protein analysis.

As two methods for the spatially resolved protein expression analysis, IMC and MIBI are promising techniques that can further refine cancer pathology and define new aspects of tissue heterogeneity with diagnostic relevance. These methods are clearly well suited to investigate how the cellular composition, protein expression and cell localisation change within a tissue in disease contexts when a well-defined set of marker proteins exist for that disease[129]. However, the reliance on antibody probes for predefined targets will limit their use in discovery settings.

1.3.2 Direct extraction from tissue

Liquid extraction surface analysis mass spectrometry (LESA-MS) was developed to address the lack of separation in MSI experiments. This sampling technique extracts molecules from the tissue or other surface into a liquid via a liquid junction[134]. In the case of proteins, these can be digested with trypsin and analysed by LC-MS/MS to provide confident peptide and protein identifications. The depth of identifications obtained is approaching that of an equivalent bulk proteomics experiment, with Ryan et al. identifying ~ 3,000 proteins from three regions of a

mouse pup tissue section[135]. However, LESA-MS has limited resolution compared to MALDI and DESI-MSI and is typically performed with resolutions of ≥ 500 µm. Another microextraction technique used to analyse proteins in tissues sections involves placing a hydrogel polymer that contains trypsin onto the tissue, digesting the proteins *in-situ* and extracting the peptides for further analysis by LC-MS/MS. Over 600 proteins were identified from a region 260 µm in diameter[136]. Neither of these methods is yet capable of accurately sampling the irregular structures found within a tissue.

1.3.3 Tissue (micro)dissection

In addition to the use of LCM to isolate tumour cells from surrounding tissue discussed above, LCM-proteomics has also been used to analyse the proteome of multiple cell types and structures within the same tissue while considering spatial location. If the tissue structures are large enough, they can be sampled by manual dissection. For instance, Guo et al. [137] used multiple 1 mm punch biopsies to investigate spatial heterogeneity in benign prostate hyperplasia and adenocarcinoma of the prostate. Interestingly, they found that more proteins showed increased variability within the regions of benign tissue than in the malignant tissue, with good agreement in immunohistochemistry (IHC) of a tissue microarray cohort. The top three most variable proteins within the benign tissue were prostate-specific antigen (PSA), acid phosphatase 3 (ACPP) and desmin (DES), and all three are used in the diagnosis of prostate cancers. Guo et al. suggest that this variability could contribute to the lack of validated protein biomarkers due to some proteins intra-tissue variance being of a similar magnitude to the inter-patient variance.

Buczak et al. [138] performed proteomic analysis of hepatocellular carcinoma tumours that were sampled spatially using LCM to quantify proteins within the bulk tumour, peri-tumour, and capsule regions. They demonstrate that functionally related proteins display heterogeneous abundance that is not explained by variation at the DNA or RNA level. This heterogeneous

abundance is conserved across patients in Serpin B3 and B4, two endogenous protease inhibitors that are clinically relevant for hepatocellular carcinoma.

Piehowski et al. [139] used LCM-proteomics to sample mouse uterine tissue in a rastered grid with a resolution of 100 µm and use their ultrasensitive nanoPOTS processing platform to generate protein images of the tissue. The end format of the data generated is like that of the mass spectrometry imaging techniques described above: an x, y grid of peptide and protein abundance, which can be co-registered to the original spatial coordinates. The excellent sensitivity of the current state-of-the-art, custom nanoPOTS robotic nanolitre-scale sample processing technique allowed for the quantification of over 2,000 proteins. The protein intensities were mapped back to the spatial location in the tissue to generate protein expression maps. The mouse uterus during early pregnancy has already been used as a model system for other MSI techniques, and their proteomic findings correlate well with MSI of lipids in the same system.

These studies highlight the need to quantify the spatial variability within tumour and normal tissues. Many diagnostic and research procedures are based on biopsy sub-samples of the tumours, and the heterogeneous spatial expression of proteins within these tumours could contribute to a dimension of inter-patient variability. A common feature of the studies discussed above is the reliance on known pathology to guide the microdissection sampling. Piehowski et al. do not rely on known pathology to guide their sampling, but they use the dominant cell type per pixel to assign pixels to groups for statistical analysis. The spatial relationship of samples to one another is not fully utilised, and so heterogeneity within these structures would be missed (unless all maps were manually inspected).

1.4 Spatial Statistics

Spatial statistics involves applying statistical methods to data with associated location information attached to extract meaning from that spatial data. One of the main applications

for these methods is in Geography, where data is inherently spatial. A fundamental assumption of traditional statistical methods is that of the independence of observations. The field of spatial statistics is based on observations that are not necessarily independent, assuming that features and processes near each other are associated somehow. This fundamental concept forms the First Law of Geography: "everything is related to everything else, but near things are more related than distant things"[140].

While a typical null hypothesis of a traditional statistical test might be that there is no difference in some distribution of measurements from multiple samples populations, a typical null for a spatial statistical method is that observed spatial data is a result of a completely random spatial process – a feature termed complete spatial randomness (CSR)[141]. Spatial statistical methods attempt to describe how unusual a spatial pattern is when compared to this standard. A CSR process has two assumptions: that each point or area has an equal probability of having the given property and that there are no dependencies between locations[142]. Analogies for protein expression in tissue would be that each region of tissue has an equal probability of expressing a protein, and the expression of that protein within one region does not affect the probability of it being expressed in another.

1.4.1 Spatial autocorrelation

Many measures and indicators of spatial relationships have been developed and used mainly in geographical, ecological, and economic studies of how features vary in space[141]. One such measure is spatial autocorrelation. As correlation refers to a relationship of observations between two variables, autocorrelation refers to a relationship between observations of a single variable[143]. Spatial autocorrelation is where the source of this autocorrelated behaviour is the spatial relationship between observations[143]. The presence and strength of a spatially autocorrelated relationship can be defined. Positive spatial autocorrelation would indicate that observations with similar values are likely to be located near each other. Negative spatial autocorrelation would indicate that similar values are likely to be dispersed and not found

near each other – or instead, dissimilar values are likely to be found near each other(Figure 5A). More extreme measures of spatial autocorrelation indicate that values are more dispersed or more clustered. In the cases where a variable shows no spatial autocorrelation, this indicates that there is no underlying process causing the spatial distribution of measured values.

Indices of spatial autocorrelation measure the dependence between values of a variable in different spatial locations. Moran's index (Moran's *I*) is one example of a measure of spatial autocorrelation[144]. Values of *I* lie between -1 and +1, where positive values indicate that areas close in space tend to have similar values, negative values indicate that areas close in space tend to have different values, and zero indicates data are randomly distributed through space (Figure 5B). As with other measures of spatial autocorrelation, Moran's *I* has been widely applied in the fields of geography, environmental sciences and ecology[145]. Moran's *I* has been used previously in studies investigating spatial distribution in radiology and biomolecule imaging in tissues[146–155].

The formula for Moran's *I* takes a form similar to a correlation coefficient and is calculated as follows:

$$I = \frac{N}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij}\left(x_i - \overline{x}\right)\left(x_j - \overline{x}\right)}{\sum_i \left(x_i - \overline{x}\right)^2}$$

Where $\sum_i \sum_j w_{ij}$ is the sum of the spatial weights matrix which encodes the pairwise spatial relationship between each observation, N is the number of observations, x_i is the value of an observation, x_j is the neighbour of x_i being considered and $\bar{x}$ is the mean of all observations. So, for regions that are neighbours, take the deviation of one region i and one region j from the overall mean and multiply together. Where regions i and j are both either larger or smaller than the overall mean, the result of this multiplication will be positive. Where one of

regions i and j is larger and one is smaller than the mean, the result of this multiplication will be negative. Each of these results is summed (if they are from a neighbouring pair), so each positive result from similar values increases the value of I and each negative result from disparate values decreases the value of I. The $\sum_i (x_i - \bar{x})^2$ and $N/\sum_i \sum_j w_{ij}$ act to standardise the calculation, bounding I to between -1 and 1[145,156].

Once a value for Moran's *I* has been calculated, its statistical significance can be determined by using a permutation test (Figure 5). Here, this permutation test considers the values of data points as fixed variables and instead randomly reassigns the spatial locations of each observed data point. As is often the case with permutations, the total number of possible rearrangements grows too quickly for all possible rearrangements to be generated. Therefore, this set of total rearrangements is randomly sampled and I is calculated for each sampled rearrangement giving a distribution to which the observed value of I can be compared. This generates a *p*-value where *p* is the proportion of permuted statistics that are more extreme than the observed value of I[141,156].

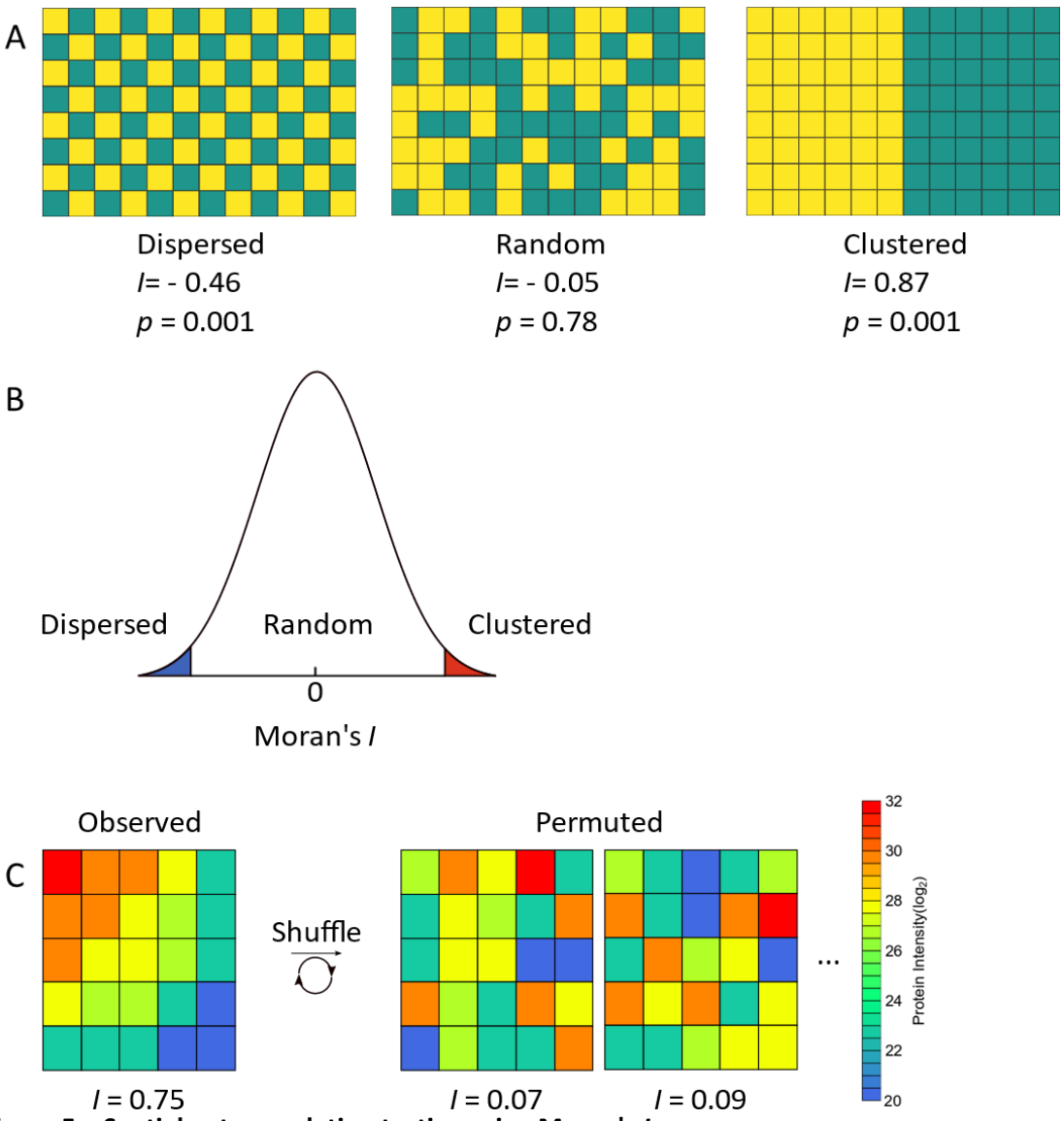

Figure 5 – Spatial autocorrelation testing using Moran's *I*

(A) Idealised example distributions of a binary variable into three spatial behaviours. Dispersed: a values neighbours are dissimilar to it. Random: levels of the variable are randomly distributed in space and equally likely to have similar and dissimilar neighbours. Clustered: levels of the variable are clustered in space and more likely to have similar neighbours. The value of Moran's *I* and associated *p*-values are shown below the example. (B) To determine significance, observed values of *I* are compared to a permuted (see C) distribution to determine significance thresholds. (C) To calculate the *p*-values in (B), the observed data is permuted randomly in space multiple times, and *I* calculated for each permutation for comparison with the observed *I*.

1.5 Atypical teratoid/rhabdoid tumours

Atypical teratoid/rhabdoid tumours (ATRT) are a class of malignant rhabdoid tumours (MRT).

MRTs are highly malignant tumours that occur throughout the body and are termed ATRT

when occurring within the central nervous system (CNS)[157]. ATRT, and all MRTs, are

characterised by loss-of-function alterations in *SMARCB1* (> 95 %) or loss-of-function alterations in *SMARCA4*[158–161]. ATRTs are classified as grade IV tumours on the WHO scale and are most commonly found within the cerebellum[162].

1.5.1 Clinical characteristics of ATRT

ATRT histology shows diverse morphology, with the epithelial, mesenchymal, neuroectodermal features – the origin of the 'teratoid' part of the name due to similarity to teratomas – and rhabdoid cells[163]. Rhabdoid cells are a hallmark of all rhabdoid tumours, with strongly eosinophilic cytoplasm, large and eccentrically located nuclei with a prominent nucleolus[164]. However, the rhabdoid cells need not account for many cells present in an ATRT tumour[165]. ATRT tumours show positive staining for vimentin, epithelial membrane antigen, cytokeratin, and smooth muscle actin. Most ATRT tumours are large and display regions of haemorrhage and necrosis[166].

The vast majority of ATRT cases are diagnosed before the age of three, with a median diagnosis age of approximately 18 months, with a median survival time of approximately 17 months[167–169]. ATRT are rare, comprising 1 – 2 % of all paediatric CNS tumours, but are responsible for a significant amount of malignant embryonal tumours[169,170].

Currently, there is no standard of therapy for ATRT, likely due to its rarity and the lack of *in vitro* and *in vivo* models[170]. Several retrospective studies have reported one-year survival estimates at less than 50 %, highlighting its high malignancy[167,168,171]. Metastatic status, tumour location and age at diagnosis are all prognostic factors. Worse outcomes are observed in the cases presenting with metastases at diagnosis[171]. Infratentorial tumours (tumours within the cerebellum and brainstem) are associated with worse outcomes when compared to supratentorial tumours (the supratentorial regions are cerebral hemispheres, thalamus, and hypothalamus)[167]. Another prognostic factor for ATRT outcome is the extent of tumour resection if surgery is performed, with gross total resection correlating with improved survival[172].

1.5.2 Treatment of ATRT

Due to the young age of most ATRT patients, radiotherapy is either not performed, delayed, or used at a lower dose to avoid the neurocognitive toxicity seen in young patients that receive radiotherapy. There is mixed evidence for the benefit of radiotherapy in the treatment of ATRT. Some studies have shown no increase in survival in patients who had received radiotherapy treatment, and others have shown increased survival in those patients receiving radiotherapy[164,173].

A now common chemotherapy-based treatment is high-dose chemotherapy, often administered as a combination of two or more of cisplatin, cyclophosphamide, vincristine, and methotrexate. This treatment is particularly attractive for patients younger than three years old as a strategy to avoid or delay radiotherapy[173]. Several studies have reported increased survival of patients receiving high dose chemotherapy[174–176]. However, there is still no consensus because of the lack of clinical trials with large cohorts.

1.5.3 Molecular characteristics of ATRT

1.5.3.1 SWI/SNF complex in ATRT

The characterising molecular feature of ATRT is the loss-of-function of *SMARCB1* (INI1), a gene that codes for a subunit of the SWI/SNF protein complex. The SWI/SNF complex is involved in the epigenetic regulation of gene expression through chromatin remodelling and is mutated in approximately 20 % of cancers[177]. In < 5 % of cases, loss-of-function of another SWI/SNF complex member, *SMARCA4* (BRG1), is present instead[160]. Other than these two characteristic alterations, ATRTs have stable genomes with no other common mutations[178]. Despite the simplistic nature of their genomic alterations, ATRTs still display a range of clinical heterogeneity[102]. This highlights the extent of how epigenetic alteration through the disruption of the SWI/SNF complex is capable of driving ATRT progression.

Knockout of *Smarcb1* in 6–10-day-old embryos by Han et al. resulted in the development of tumours that resemble ATRT based on the anatomical, morphological, and molecular

profiles[179]. Inactivation of *Smarcb1* earlier than day 6 or at later time points resulted in embryonic lethality and liver toxicity or the development of lymphomas. The loss of *Snf5* (the *Drosophila* homologue of *SMARCB1*) results in an increased expression of the *EZH2* histone methyltransferase, the catalytic subunit of the Polycomb repressor complex 2 (PRC2)[180]. PRC2 suppresses gene expression of its targets by trimethylating histone 3 proteins at lysine 27 within the promotor region. EZH2 overexpression has been shown to promote cell survival and proliferation[181]. A double knockout of *Smarcb1* and *Ezh2* in mice was shown to prevent tumorigenesis, leading to clinical trials into the effectiveness of *EZH2* inhibitors in ATRT.

1.5.3.2 Molecular subtypes

Epigenetic and transcriptomic studies have recently demonstrated the presence of multiple molecular subtypes of ATRT. Torchia et al. identified two transcriptional subgroups in a study of 259 patients: A subgroup showing expression of ASCL1 and mainly located supratentorialy had higher survival at five years and ASCL1-negative tumours showing lower five-year survival in patients who received either radiotherapy and chemotherapy or chemotherapy alone – these were termed group 1 and group 2 respectively[182]. Torchia et al. and Johann et al. both reported the presence of three molecular subtypes from genetic, epigenetic, and transcriptomic data showing the subtypes have distinct methylation patterns with their own potentially targetable pathways[178,183]. The three subtypes were termed group 1 (ATRT-SHH), group 2a (ATRT-TYR) and group 2B (ATRT-MYC). The group 1 subtype have neurogenic profiles and show an increase in *ASCL1* and Notch pathway expression, have a closed chromatin landscape with generally decreased gene expression and were predominantly supratentorialy located. The group 2a/b subtypes show increased Bone morphogenetic protein (BMP) signalling genes and mesenchymal differentiation. Group 2a showed enrichment of retinol and tyrosine metabolism gene expression along with an open chromatin landscape and were located predominantly infratentorialy located. Group 2b showed upregulation of *MYC* and

HOX gene expression, a closed chromatin landscape and were found in more heterogeneous locations[102].

1.5.3.3 Potential future therapies

The knowledge gained from molecular subtyping and cell line studies has identified a range of epigenetic and signalling inhibitor targets for evaluation in clinical trials. After their subtype analysis, Torchia et al. tested several compounds based on their molecular findings[178]. The γ-secretase (the protease responsible for cleavage and activation of Notch) inhibitor DAPT was found to inhibit ATRT cell lines with a group 1 phenotype in a dose-dependent response. Group two cell lines were found to be sensitive to dorsomorphin, an inhibitor of BMP signalling, in a dose-dependent response. The histone deacetylase (HDAC) inhibitor LAQ824i reduced growth when tested in both group 1 and group 2 ATRT cell lines. Another HDAC inhibitor, SNDX-275, has been shown to induce apoptosis in two ATRT cell lines[184].

1.6 Thesis aims

There is a clear need for techniques to investigate protein expression heterogeneity within biological tissue structures for many proteins simultaneously. The focus of this thesis was the development of methods for spatially resolved, mass spectrometry-based proteomics. This was performed in the context of human brain neurons and within a rare paediatric brain tumour. The specific aims of the project were:

1. Develop an optimised proteomic sample preparation workflow to detect protein abundance differences with cell-type resolution in the human brain.

2. To extend this optimised method to an unbiased sampling approach that retains the spatial context of isolated cells from the tumour and adjacent tissue.

3. To develop data analysis workflows which use this retained spatial context, without the need for prior histopathology-based tissue annotation, to determine spatially variable proteins within tumour tissue.

2 Development of a sensitive, quantitative method for LCM proteomics

Reproduced in part with permission from ([185]) Davis et al., Journal of Proteome Research 2019 18 (4), 1787-1795 DOI: 10.1021/acs.jproteome.8b00981 Copyright 2019 American Chemical Society.

2.1 Introduction

The purpose of the work presented in this chapter was to develop a workflow for sensitive, spatially resolved proteomics of human brain tissue using laser capture microdissection. Tissue heterogeneity is a generally overlooked factor in clinical proteomics studies, as large amounts of input material are often needed. Laser capture microdissection (LCM) is well-suited as a method to investigate phenotypic and spatial differences within tissues. LCM allows for the precise isolation and extraction of individual cells or features of interest from surrounding tissue[186]. This ability is critical for the avoidance of the analyte averaging issues of bulk samples, as analysis on or approaching the single-cell level allows for the interrogation of cell subpopulations. It is complementary to flow cytometry-based methods as cell type-unique antibody markers are not necessarily required. Also, LCM can more easily handle cells that are too large or complex for cell sorting, such as large neurons in the brain. LCM has been widely used for proteomic analysis of a range of tissues[138,187–189]. The proteomic analysis of small subpopulations requires either large amounts of input material to harvest the desired cells or features, or sensitive sample preparation and analysis workflows.

Advances have been made in all steps of the proteomic workflow to facilitate the analysis of trace sample amounts. For example, the development of Single-pot, solid-phase-enhanced sample preparation (SP3), a workflow for the processing of trace samples in a single reaction vessel, which uses protein/peptide binding to paramagnetic beads to remove detergent, digest protein, and clean-up peptides[22]. More recently, Zhu et al. (2018b) described a semiautomatic nanodroplet-based platform capable of performing protein digestion in <200 nL of volume (nanoPOTS). In combination with identification transfer from a larger sample using

MaxQuant's "match between runs" feature, ~2700 proteins could be quantified for samples containing 10–14 HeLa cells. The same group also demonstrated the nanoPOTS platform's ability to analyse single cells collected by fluorescence-activated cell sorting (FACS), and the ability to detect ~670 proteins from a single HeLa cell[191]. Both studies used a custom platform with a narrower C18 column and lower flow rate than is typically used in proteomics experiments (30 µm i.d. and 60 nL/min respectively), resulting in gains in sensitivity [192]. This custom robotics platform is inaccessible to most proteomics laboratories, which limits its use outside the group where it was developed. In addition, the low flow rates used with narrow-bore LC columns do not allow for the analysis of many samples per day due to a large time delay between the formation of the gradient in the pump and it reaching the column. Furthermore, the original work does not use an autosampler to introduce samples into the system, although this has been addressed in further work through the use of a prototype autosampler[193]. The nanoPOTS platform currently represents the state-of-the-art, but more accessible methods for sample preparation of low input amounts are required.

As a test case for the here developed method, two easily identifiable neuron classes were isolated by LCM for proteomic analysis. These cells are the Betz cells of the motor cortex and the Purkinje cells of the cerebellum. Both cell types are the primary efferent cells of their respective layers and are among the largest human brain cells[194,195]. Purkinje cells are well studied, with several well-known proteins expressed that could act as internal controls that would be expected to be identified within the Purkinje cell proteome. Identifying the defining proteomic signature of these (and other) cells is relevant not only for basic neuroscience research (for example, through the unbiased identification of cell-type-specific protein markers for the generation of antibody-based investigative tools), but also for a better understanding of neurological diseases.

It was hypothesised that these two morphologically, topographically, and functionally distinct neuronal cell types would serve as a suitable proof of concept to test the developed method for spatially resolved proteomics using laser capture microdissection. Furthermore, there are

no known Betz cell-specific markers, so this test case also presents the opportunity to gain further insights into the proteome of these understudied cells.

2.2 Materials and Methods

Reagents were purchased from Sigma Aldrich unless otherwise specified. Solvents for LC-MS use were LC-MS grade and purchased from Millipore. Formic acid was purchased from Fisher Scientific.

2.2.1 Tissue retrieval and processing

Brain tissue was retrieved 3 days *post-mortem* by the Oxford Brain Bank; a research ethics committee (REC) approved, and Human Tissue Authority-regulated research tissue bank (REC reference 15/SC/0639). After retrieval, the brain was divided into the two hemispheres. One was fixed in 10 % Neutral Buffered Formaldehyde, the other hemisphere was sliced in the coronal plane and individual slices of motor cortex or cerebellum were immediately placed into liquid nitrogen vapour and stored at -80 °C. This method substantially reduces freezing artefacts resulting in frozen tissue sections with excellent cytoarchitecture (Figure 6), which is critical for accurate LCM.

Relevant tissue blocks of the frozen cerebellum and motor cortex were acclimatised to -20 °C and mounted onto a cryostat block using OCT Compound (Cell Path, ARG1180). Careful consideration was taken to ensure cut sections were not contaminated with OCT. Sections were cut at 10 µm and mounted onto UV irradiated (254 nm, 30 minutes) 1.0 PEN membrane slides (Zeiss) at -18 °C. Sections were then air-dried for several minutes and placed onto a Shandon Linistain for automated H&E staining. Briefly, sections were fixed in 70 % denatured alcohol, hydrated, stained with Harris' Haematoxylin, incubated in 0.4 % acid alcohol, placed in Scot's tap water, and stained with Eosin containing 0.25 % acetic acid with regular washing steps in between. Stained sections were then dehydrated in increasing concentrations of denatured alcohol and air-dried without coverslips and stored at -80 °C until processing by laser-capture microdissection.

2.2.2 Laser-capture microdissection

Areas of tissue and multiple, individual sections of Purkinje and Betz cells were isolated from the prepared cerebellum and motor cortex slides using a laser-capture microscope equipped with laser pressure catapulting (PALM Microbeam, Zeiss). All annotations of areas/cells for capture were performed manually on the microscope. Purkinje and Betz cells were annotated based on histological landmarks and annotated according to their size, location, and morphology. Cutting and catapulting of areas/cells of interest were performed along annotations automatically using the 40x objective lens and the following settings in the control software for cutting: Energy 43, Focus 55; and catapulting Energy 20, Focus -15. After isolation, samples were immediately placed in dry ice and later stored at -80 °C.

2.2.3 Proteomic sample processing

2.2.3.1 Tissue lysis

Protein was extracted from tissue collected in the LCM cap by adding 10 µL of lysis buffer followed by both brief centrifugation and incubation at room temperature for 30 minutes, or by omitting the centrifugation step and incubating in an inverted position for 30 minutes at room temperature.

2.2.3.2 In-cap digestion

Protein extracts in the LCM cap were diluted to either 1 M Urea or 10 % Trifluoroethanol (TFE) with 50 mM ammonium bicarbonate, 25 ng of trypsin (V5111, Promega) was added, and samples incubated at 37 °C overnight – all steps taking place in the LCM cap. Peptides were acidified to 1 % formic acid (v/v) (Fisher Scientific) and purified using C18 solid-phase extraction cartridges (SOLA HRP, Thermo Scientific). Peptides were dried to completeness under vacuum, resuspended in 6 µL 2 % acetonitrile, 0.1 % trifluoroacetic acid, and stored at -20 °C until analysis.

2.2.3.3 In-solution digestion

Protein lysates within the LCM cap (either after collecting dry or directly into the buffer) were reduced and alkylated in-cap with 5 mM DTT and 20 mM iodoacetamide (final concentrations,

see Table 1). Lysates were then transferred to 200 µL PCR tubes. Proteins were cleaned-up before digestion by precipitation with methanol-chloroform [17]. The protein pellet was resuspended in 6M urea 0.1 M tris pH 7.8; this was then diluted to 1 M urea with ultrapure water, 25 ng trypsin added and incubated at 37 °C overnight. Peptides were acidified to 1 % formic acid (v/v) and purified using C18 solid-phase extraction cartridges (SOLA HRP, Thermo Scientific). Peptides were dried to completeness under vacuum, resuspended in 6 µL 2 % acetonitrile, 0.1 % trifluoroacetic acid, and stored at -20 °C until analysis.

2.2.3.4 Single-Pot Solid-Phase-enhanced Sample Preparation (SP3)

Paramagnetic beads (GE45152105050250 & GE65152105050250, Cytiva) were prepared as described [22]. Protein lysates within the LCM cap (either after collecting dry or directly into the buffer) were reduced and alkylated in-cap with 5 mM DTT and 20 mM iodoacetamide (final concentrations, see Table 1). Lysates were then transferred to 200 µL PCR tubes. 2 µL of beads were mixed with the samples. Acetonitrile was added to the samples to a final concentration of 70 % (v/v) and briefly vortexed. This mixture was incubated for 18 minutes with orbital shaking at 1,000 rpm. Beads were then immobilised on a magnet for 2 minutes and the supernatant transferred to fresh tubes containing 2 µL of fresh beads and the incubation step repeated to capture any initially unbound protein. Beads were washed twice with 70 % (v/v) ethanol in water and once with 100 % acetonitrile, all on the magnet. Beads were resuspended in 50 mM ammonium bicarbonate containing 25 ng trypsin and digested overnight at 37 °C. After digestion, the beads were resuspended by bath sonication. Acetonitrile was added to 95 % (v/v) and beads incubated at 1,000 rpm for 18 minutes to bind peptides, then beads were immobilised on the magnet for 2 minutes and the supernatant discarded. Peptides were eluted from the beads with 6 µL 2 % DMSO (Fisher Scientific). Beads were pipette mixed and then immobilised on the magnet for 5 minutes, and peptides transferred to LC-MS vials containing 600 nL 1 % trifluoroacetic acid. Vials were stored at -20 °C until analysis.

2.2.4 LC-MS/MS

Peptides were analysed by nLC-MS/MS using a Dionex Ultimate 3000 coupled on-line to an Orbitrap Fusion Lumos mass spectrometer (Thermo Scientific). A 75 µm x 500 mm C18 EASY-Spray column (Thermo Scientific) with 2 µm particles was used at a flow rate of 250 nL/min for all samples. Peptides were separated using a 60-minute linear gradient from 2 % buffer B to 35 % buffer B (A: 5 % DMSO, 0.1 % formic acid in water; B: 5 % DMSO, 0.1 % formic acid in acetonitrile). For the initial Purkinje cell experiments (Figure 7), the mass spectrometer was operated with the following settings ("Universal method" [196]): MS scans were acquired in the Orbitrap between 400 and 1500 *m/z* at a resolution of 120000 and an automatic gain control (AGC) target of $4x10^5$. Selected precursors between charge state 2+ and 7+ and above the intensity threshold of $5x10^3$ were selected for collision-induced dissociation (CID) fragmentation, with a collision energy of 35 %, an AGC target of $4x10^3$, a maximum injection time of 300 ms, a dynamic exclusion window of 60 seconds and a maximum duty cycle of 3 seconds. MS/MS spectra were acquired in the ion trap using the rapid scan mode. The option to "use all available parallelisable time" was enabled.

For the molecular layer optimisation experiments and Betz/Purkinje comparison, the following settings were used: MS1 scans were acquired as above. Selected precursors between charge states 2+ and 7+ and above the intensity threshold of 5e3 were selected for higher-energy collisional dissociation (HCD) fragmentation, with a normalised collision energy of 28 %, an AGC target of $4x10^3$, a maximum injection time of 80 ms, a dynamic exclusion window of 30 seconds and a maximum duty cycle of 1 second. MS/MS spectra were acquired in the ion trap using the rapid scan mode.

2.2.5 Data analysis

The raw data files were searched against the Uniprot human database (Retrieved 17/01/2018, 929527 sequences) using MaxQuant [57,197] version 1.6.2.10 and its built-in contaminant database using tryptic specificity and allowing 2 missed cleavages, peptide and protein

identifications are reported using the protein parsimony principles as described by Nesvizhskii and Aebersold [49]. Results were filtered to a 1 % false discovery rate at the PSM and Protein level. At least 1 unique peptide was required to identify a protein group. Cysteine carbamidomethylation was set as a fixed modification as appropriate, methionine oxidation and protein N-terminal acetylation were set as variable modifications, and the 'match between runs' option was used. All other settings were left as default. Protein quantification was performed with the MaxLFQ algorithm within MaxQuant [66]. Downstream analysis was performed in Perseus [198] and R [199].

2.2.6 Immunohistochemistry

Sections from FFPE blocks were cut at 6 μm and mounted onto superfrost microscope slides. Once air-dried and baked at 56 °C for 20 minutes, sections were deparaffinised and rehydrated with incubation in xylene, 100%, 90%, 70% denatured alcohol, and tap water. Rehydrated sections were incubated for 30 minutes in 10% hydrogen peroxide to remove endogenous peroxidases. Antigen retrieval was performed by autoclave heating (121 °C for 10 minutes) in Sodium Citrate Buffer (pH 7.6). Mouse anti-calbindin was diluted in TBS/T (1:1000) and incubated overnight at 4 °C. Staining visualisation was performed using the Aligent's Dako Envision Kit (Aligent, Cat. number k5007). Briefly, a cocktail of mouse/rabbit secondary antibodies conjugated to HRP were incubated for 1 hour at room temperature. A 2% DAB working solution was prepared with a 50x concentrated 3,3'-diaminobenzidine (DAB) stock and diluted with Dako's substrate buffer, and sections were incubated for 5 minutes then counterstained with Harris' Haematoxylin for 1 minute. This tissue sectioning and staining was kindly performed by Connor Scott

Table 1 – Protein preparation steps for combinations of lysis buffers used in SP3 and In-solution digestion

Protein solubilisation system	Collection Method	Collection Buffer	Protein recovery & reduction	Protein alkylation
6M urea	Buffer	20 µL 6M urea, 0.1 M tris	Add DTT to 5 mM. Incubate 30 mins, RT	Add IAA to 20 mM. 30 mins, RT
	Cap	—	Add 10 µL 6M urea, 0.1 M tris, 5 mM DTT. Incubate 30 mins, RT	
6M urea, 30 % acetonitrile	Buffer	20 µL 6M urea, 100 mM Tris, 30 % acetonitrile (v/v)	Add DTT to 5 mM. Incubate 30 mins, RT	
	Cap	—	Add 10 µL 6M urea, 0.1 M tris, 5 mM DTT, 30 % acetonitrile (v/v). Incubate 30 mins, RT	
50 % Trifluoroethanol	Buffer	20 µL Ultrapure water	Add 20 µL trifluoroethanol. Add DTT to 5 mM. Incubate 60 mins, 60 °C	
	Cap	—	Add 10 µL 50 % trifluoroethanol (v/v), 0.1 M tris, 5 mM DTT. Incubate 60 mins, 60 °C	
RIPA	Buffer	20 µL RIPA	Add DTT to 5 mM. Incubate 30 mins, RT	
	Cap	—	Add 10 µL RIPA (Pierce #89900), 5 mM DTT. Incubate 30 mins, RT	

2.3 Results and Discussion

The goal of the work presented here was to develop a sensitive and quantitative protocol capable of generating data at a spatial resolution. The initial work was performed on fresh-frozen material sampled from the human cerebellum due to the presence of Purkinje cells, which are easily identifiable after H&E staining due to their defined location within a monolayer and large size relative to other neurons (Figure 6).

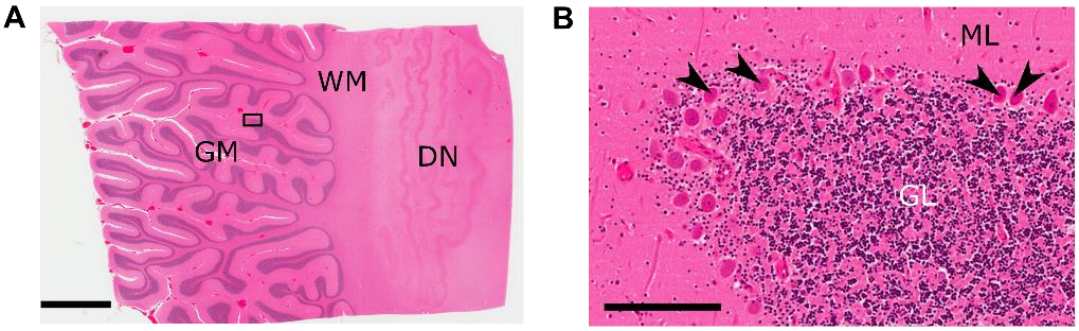

Figure 6 – Location of Purkinje cells within the cerebellum.

(A) Image of a cerebellar cortex tissue slice stained with H&E. Rectangle indicates area shown in B. GM: grey matter. WM: white matter. DN: dentate nucleus. Scale bar represents 4 mm. (B) Higher magnification of the area indicated in (A). Arrowheads indicate Purkinje cells. ML: molecular layer. GL: granular layer. Scale bar represents 200 µm. Reproduced from Davis et al. 2019

2.3.1 Defining a simple starting point for LCM proteomics

To define an approximate minimum area of tissue that could be isolated and generate enough identifications to be of use, sections of 100, 200, 400, and 800 Purkinje cell bodies were isolated from 10 µm thick H&E sections by laser-capture microdissection (LCM). The area covered by the 100 Purkinje cell body sections was approximately 60 000 μm^2, resulting in a tissue volume of approximately 600 pL, assuming a cellular protein concentration of 200-300 g/L [200]; this sample contained approximately 120-180 nanograms of protein. Using a Purkinje cell body volume of 12,500 μm^3, 600 pL represents approximately 50 Purkinje cells[201]. Samples were recovered from the LCM caps by adding 6 M urea followed by centrifugation, dilution to

1 M urea, and digested with trypsin. From these samples, 108, 181, 4221, and 6728 peptides from 73, 107, 1517, and 2051 proteins were identified, respectively (union of identifications by MS/MS and match-between-runs) (Figure 7).

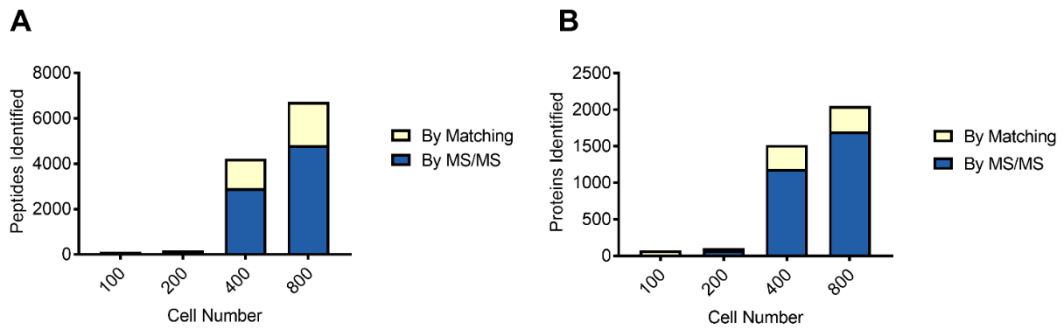

Figure 7 – Peptide and protein identifications from digest after recovering cells from the LCM cap.

As a pilot experiment, 100, 200, 400, and 800 Purkinje cells were collected using LCM and processed after centrifuging to collect the cells from the cap before digestion. 'By matching' refers to identifications matched from a cerebellar cortex sample using MaxQuant's 'match between runs'. (A) The number of peptides identified from Purkinje cells recovered from the LCM cap. (B) The number of proteins identified from Purkinje cells recovered from the LCM cap. Reproduced from Davis et al. 2019

Following the material's recovery from the LCM caps by centrifugation, some material was still visible in the LCM cap. The presence of this material, the low total ion current signal, and the number of identifications for the 100 and 200 Purkinje cell body section samples suggested that this material could still contain protein. To test this, the experiment was repeated, and the centrifugation steps were omitted. Simply digesting the proteins in-place without recovery from the LCM cap – descriptively referred to as the "In-Cap" method (Figure 8) to distinguish this approach from a more conventional in-solution digestion (ISD) method below.

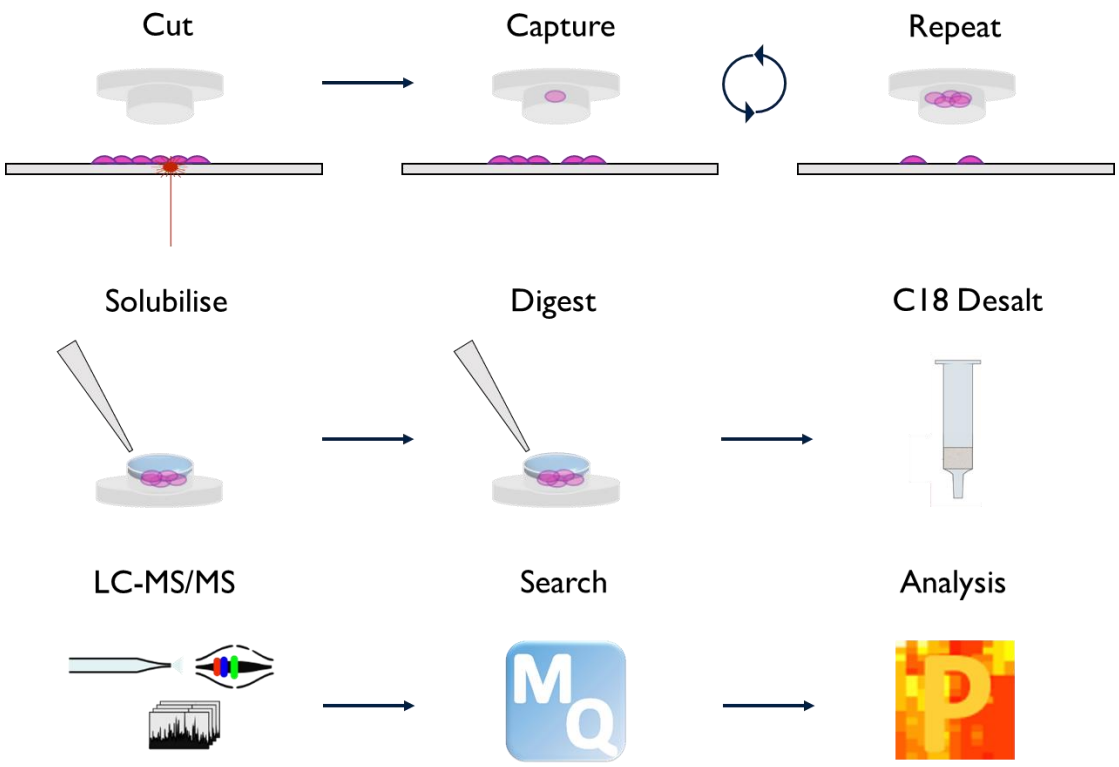

Figure 8 – In-cap digestion workflow.

Cells are repeatedly cut away from the surrounding tissue and captured into an LCM cap. Proteins are then solubilised and digested in the cap, removing the need to recover material from the cap. Peptides are then desalted and analysed by LC-MS/MS. Raw data are searched in MaxQuant and analysed in Perseus.

This modification showed an increase in the total ion current signal and identifications (Figure 9). Using this in-cap workflow, 3852, 8921, 8855, and 8858 peptides from 1497, 2224, 2360, and 2282 proteins were identified, respectively (union of identifications by MS/MS and match-between-runs) (Figure 10). Within the proteins identified using this method, 15 proteins detected proteins are annotated as being involved in Purkinje cell development or function, along with almost 590 proteins involved in general neurological cell function. The digestion efficiency was good; 75% of peptides had no missed cleavages, 22 % had one missed cleavage, and 3 % had two missed cleavages. For simplification, the typical steps of reduction and alkylation were omitted, which is noticeable in the data. Less than 1 % of peptides detected contained a cysteine residue, compared to 8 % of peptides detected in further experiments

described below, where the reduction and alkylation steps were performed. There was little to no increase in protein and peptide identifications when over 200 Purkinje cell body sections were sampled. As the peptide and protein identifications obtained from the 100 Purkinje cell body section sample were below the observed plateau, this tissue amount was used for further optimisation experiments.

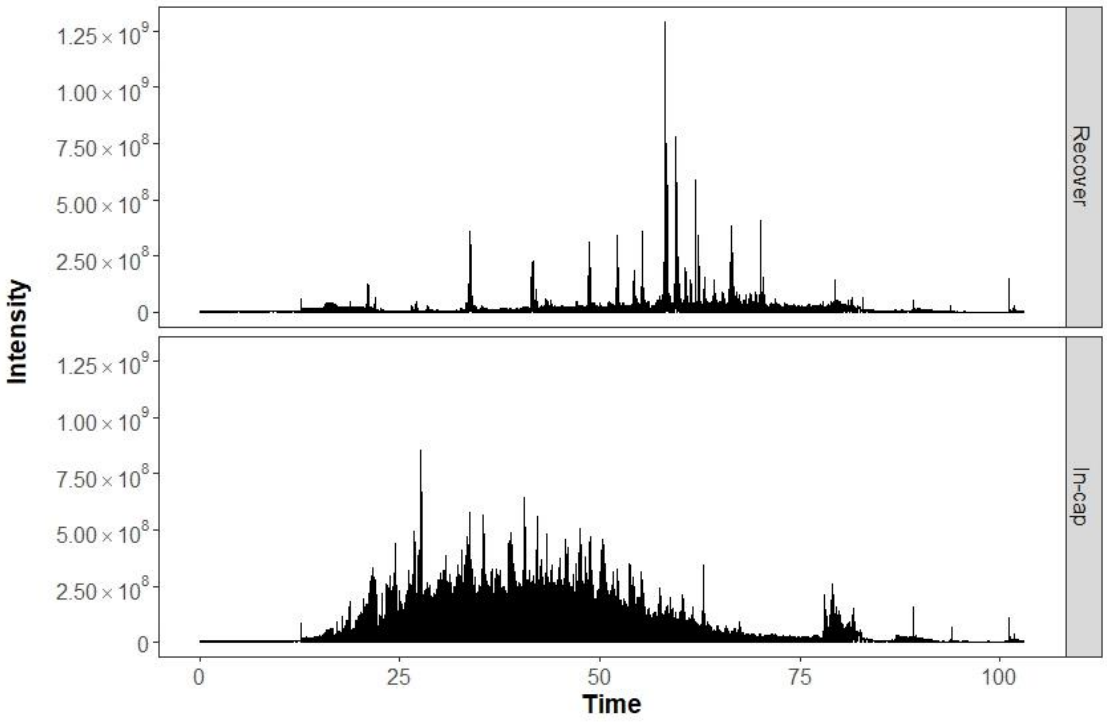

Figure 9 – Total ion chromatogram showing In-cap digestion improvement

Total ion chromatogram of a 200 Purkinje cell sample where the material was attempted to be recovered from the LCM cap (Top) and where the material was digested *in-situ* in the LCM cap (Bottom).

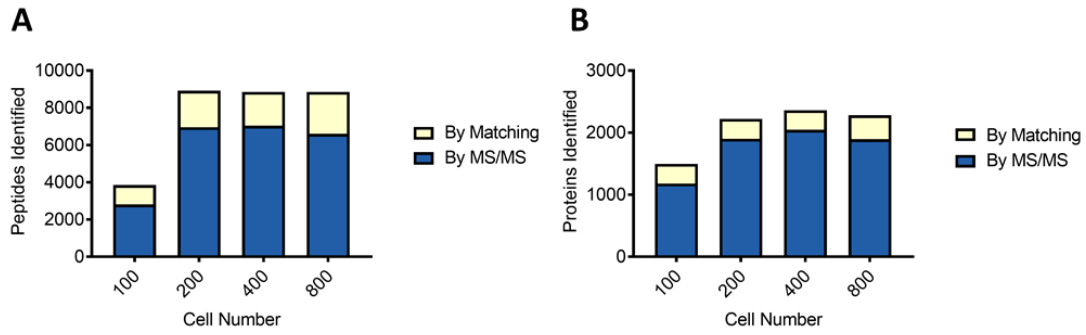

Figure 10 – Peptide and protein identifications from the In-cap workflow.

100, 200, 400, and 800 Purkinje cells were collected using LCM and digested directly in the LCM cap. 'By matching' refers to identifications matched from a cerebellar cortex sample using MaxQuant's 'match between runs'. (A) The number of peptides identified from Purkinje cells digested in the LCM cap. (B) The number of proteins identified from Purkinje cells digested in the LCM cap. Reproduced from Davis et al. 2019

2.3.2 Towards an optimised method for LCM proteomics

After demonstrating that a useful number of proteins can be identified from as little as 100 Purkinje cell body sections with a simple method, a series of optimisation experiments were performed to investigate the effects of the following parameters: (i) collection method, (ii) lysis buffer and (iii) digestion protocol. Given the combinations of factors, these experiments were performed on the cerebellum's molecular layer by isolating 60,000 µm^2 of tissue (approximately equivalent to the 100 Purkinje cell body sections used above). This removed the need for manual location and annotation of thousands of individual cells, dramatically reducing the amount of time needed to perform the LCM on the fresh-frozen tissue and allowed all samples to be collected from the same tissue slice. Figure 11 shows a schematic overview of this experiment. The three digestion methods tested were: (i) a scaled-down version of our lab's in-solution digestion protocol, (ii) a modified version of SP3 [202], and (iii) the in-cap method used to generate the Purkinje cell data. In combination with these digestion methods, four different collection and lysis buffers were tested: (i) 6 M urea, (ii) 6 M urea in 30 % acetonitrile [203], (iii) 50 % trifluoroethanol (TFE) [204], and (iv) RIPA buffer. Finally, each lysis buffer/method combination was performed by either collecting material into a dry, polymer-

filled LCM cap with subsequent extraction or by collection directly into the buffer. The exception is that RIPA buffer was not used in combination with the in-cap method. This method does not use a detergent removal step, and the detergents present in RIPA buffer interfere with peptide elution from C18 liquid chromatography columns.

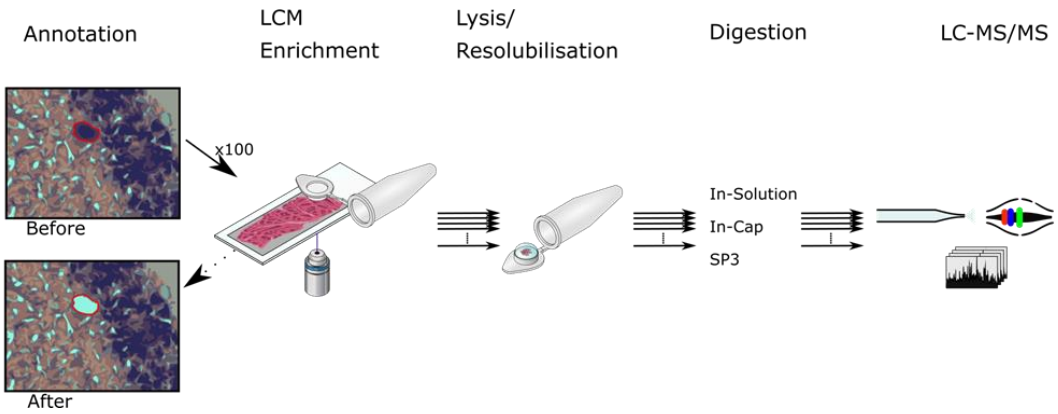

Figure 11 – Schematic overview of the LCM workflow.

One hundred disks of the cerebellar molecular layer were collected per sample for a total area of 60,000 µm^2. Different collection methods, lysis buffers, and digestion protocols were tested. Reproduced from Davis et al. 2019

The various sample preparation approaches performed quite differently for this tissue type and amount, with the detection of between 38-6399 peptides corresponding to 17-1521 proteins (mean, n = 3) (Figure 12). In the context of peptide and protein identifications, the combination of sample collection directly into 50 % TFE and processing with SP3 (Buffer-TFE-SP3) generated 6399 peptide and 1521 protein identifications. The methods generating the next highest numbers of identifications were Buffer-RIPA-SP3 (1316 proteins), Cap-RIPA-SP3 (1268 proteins), and Buffer-RIPA-In-Solution (1260 proteins). The overlap in protein identifications between the two methods with the most identifications (Buffer-TFE-SP3 and Buffer-RIPA-SP3) was 74% when considering proteins identified in at least two replicates per condition (Figure 13).

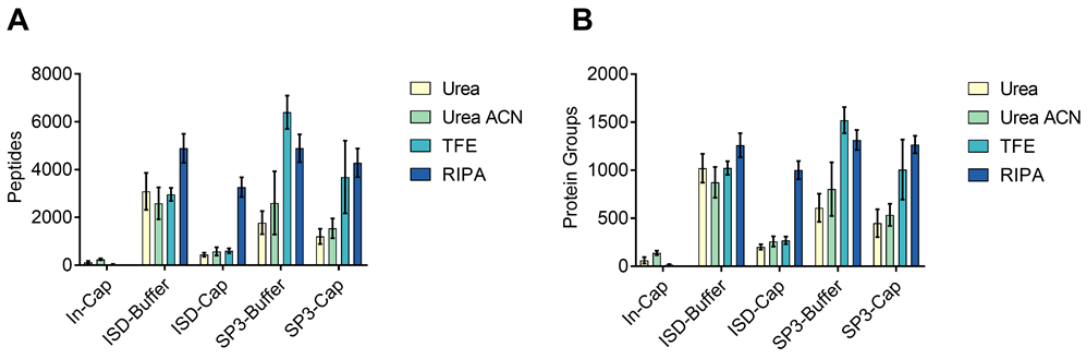

Figure 12 – Comparison of protein identifications of methods used for protein retrieval and digestion.

To determine an optimal method of material collection, lysis, and digestion, 60 000 µm3 of the molecular layer (the area of approximately 100 Purkinje cells) was collected to test different combinations of collection methods, lysis buffers, and digestion methods. (A) Mean number of peptides identified in each combination. (B) Mean number of proteins identified in each combination. All bars are mean values (n = 3), and error bars represent standard deviation. In-Cap refers to the in-situ digestion method. Cap and Buffer refer to collecting material onto both a dry LCM cap and recovering in the indicated lysis buffer, or collecting directly into the indicated lysis buffer. Reproduced from Davis et al. 2019

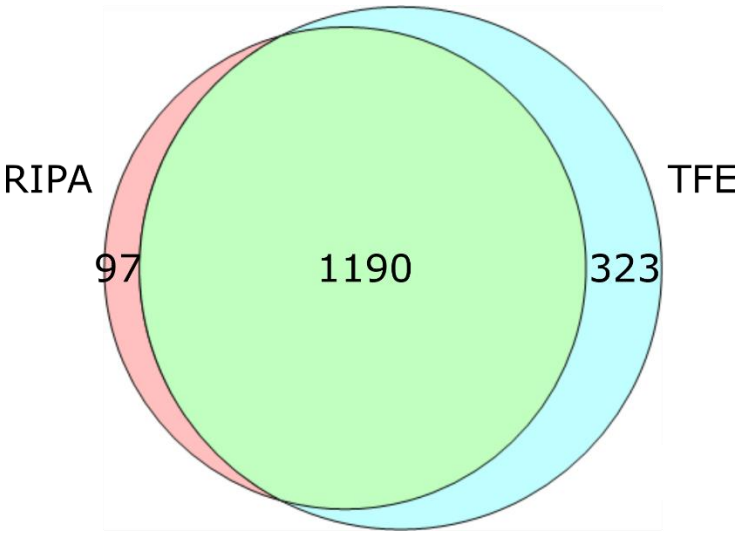

Figure 13 – Identification overlap of proteins detected in RIPA and TFE.

The overlap in protein identifications between the two methods with the most identifications (Buffer-TFE-SP3 and Buffer-RIPA-SP3) considering proteins identified in at least two replicates per condition. Reproduced from Davis et al. 2019

To further investigate the larger proportion (and number) of unique identifications in the TFE sample, these proteins' properties were investigated. Figure 14 shows that RIPA's protein

recovery was slightly biased toward recovering proteins of a lower molecular weight. In contrast, the distribution of proteins identified in TFE samples more closely resembles the distribution of proteins in the reviewed SwissProt database (20 336 entries). Next, the isoelectric points of the uniquely identified proteins were compared to the SwissProt database. Proteins uniquely identified in TFE samples tended to be more acidic than those identified by RIPA alone, and RIPA uniquely identified more proteins with a basic isoelectric point (Figure 15). The grand average of hydropathy (GRAVY) [205] index is a measure of an amino acid sequence's hydropathy, where more negative values indicate more hydrophilic sequences. These indices were calculated for each of the proteins uniquely identified in either the TFE or RIPA samples and their distributions compared to all proteins in the Swiss-Prot database (Figure 16). TFE shows a slight bias toward hydrophilic proteins (median GRAVY value = −0.41), and RIPA shows a slight bias toward more hydrophobic proteins (median GRAVY value = −0.32).

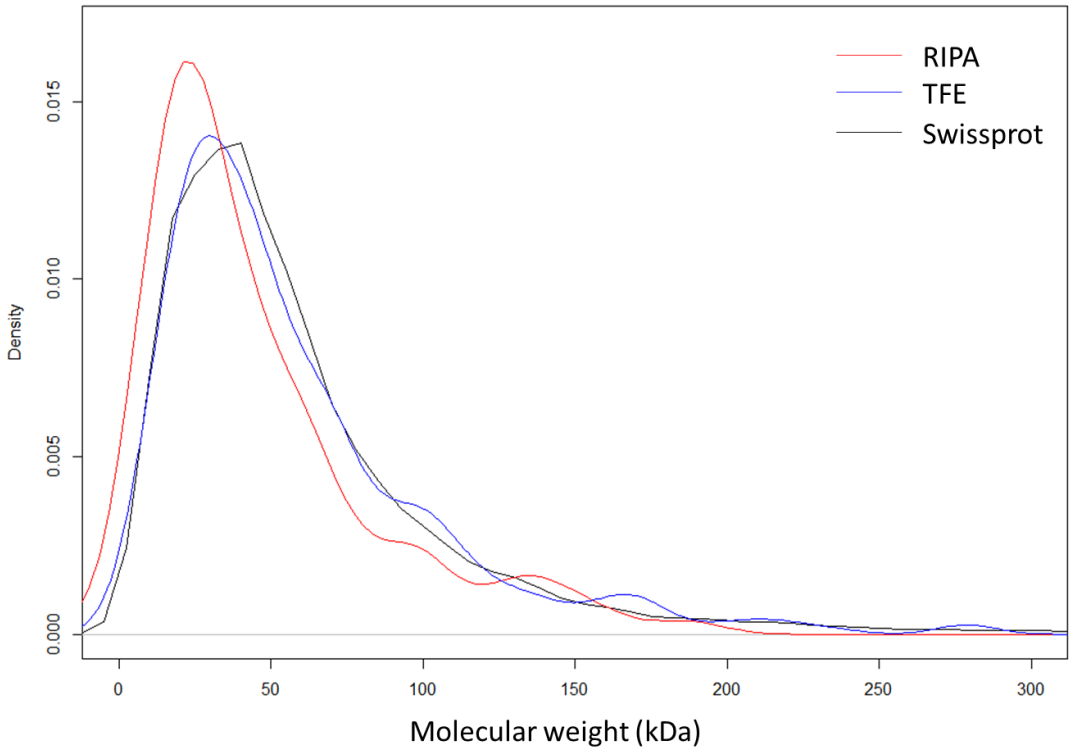

Figure 14 – Protein molecular weight distributions

Kernel density estimates of the distribution of the molecular weight of proteins identified in samples processed with RIPA and TFE lysis buffers compared to that of all human proteins in the SwissProt database. Reproduced from Davis et al. 2019

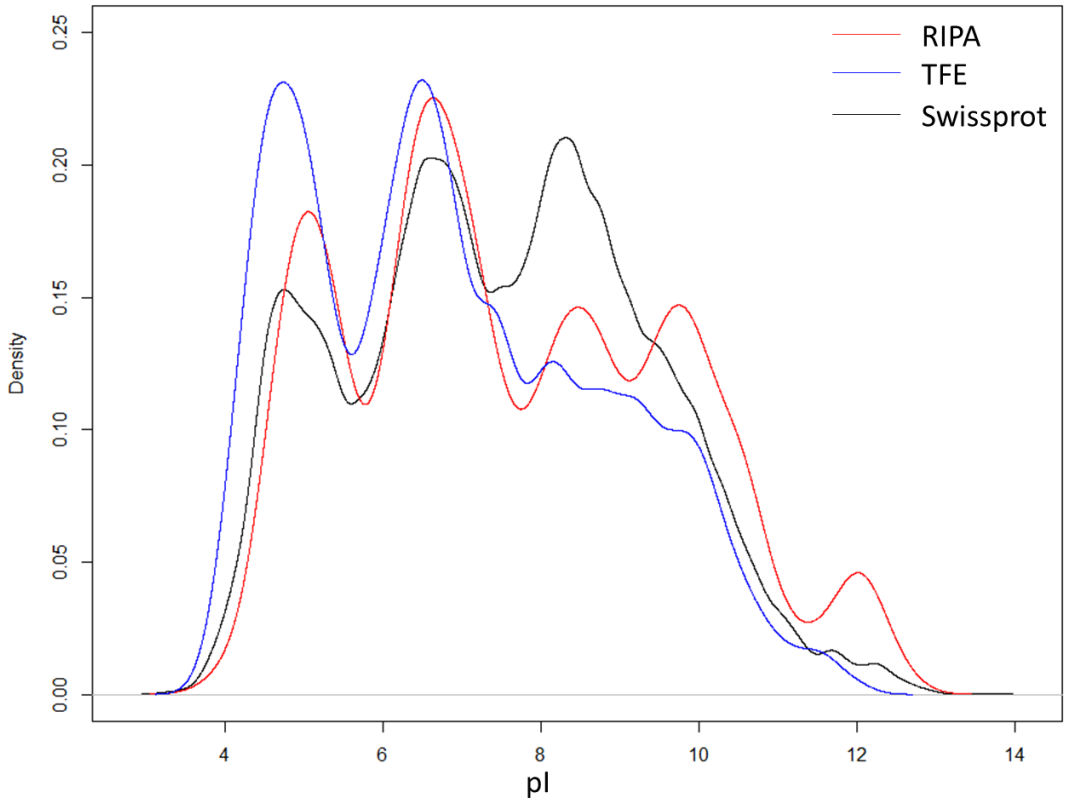

Figure 15 – Protein isoelectric point distributions

Kernel density estimates of the distribution of the isoelectric point of proteins identified in samples processed with RIPA and TFE lysis buffers compared to that of all human proteins in the SwissProt database. Reproduced from Davis et al. 2019

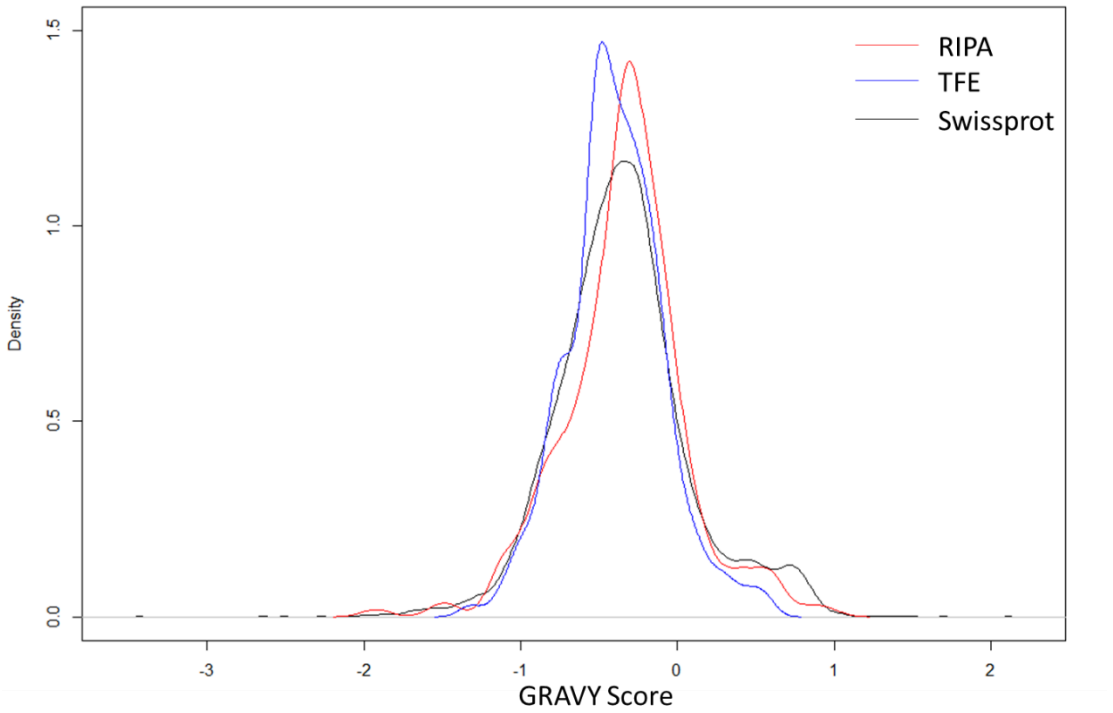

Figure 16 – Protein GRAVY score distributions

Kernel density estimates of the distribution of GRAVY hydropathy of proteins identified in samples processed with RIPA and TFE lysis buffers compared to that of all human proteins in the SwissProt database. Reproduced from Davis et al. 2019

In general, collecting samples directly into a liquid performed better than collecting directly onto the LCM polymer cap and recovering protein afterward. When 50 % TFE was used to recover proteins from the LCM cap, the number of identifications decreased and became more variable than collecting into the buffer, likely still due to incomplete recovery of protein from the cap as indicated by a reduction in the summed intensity of identified peptides (Figure 17). The SP3 method outperformed the in-cap method in all cases, whereas the in-solution method only outperforms the in-cap method when the sample is collected into liquid. The in-cap method could not be performed with RIPA buffer because it does not contain a detergent removal step.

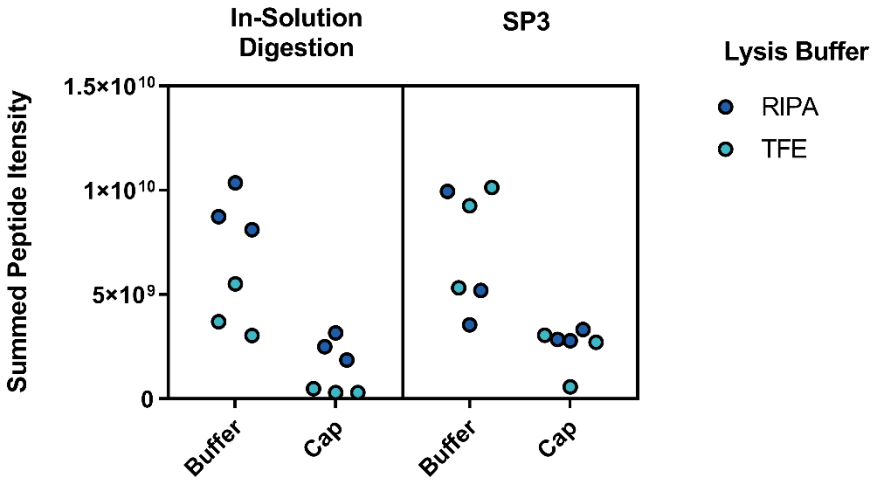

Figure 17 – Summed peptide intensity of RIPA and TFE samples

Summed MaxQuant peptide intensities of molecular layer optimisation experiments processed with RIPA and TFE lysis buffers using either the in-solution digestion or SP3 digestion methods after collecting into either lysis buffer or directly onto the LCM cap. Reproduced from Davis et al. 2019

As well as identification level performance, we also investigated the quantitative performance of the methods tested. As would be expected, the number of proteins quantifiable using MaxQuant's MaxLFQ algorithm follows the same pattern as the number of protein identifications across each method combination (Figure 18A). The highest number of quantified proteins was 958 (median, n = 3) in the Buffer-TFE-SP3 samples. Samples processed with either the in-solution digestion protocol or SP3 combined with TFE or RIPA buffer all had high Pearson correlation coefficients within replicates (n = 3)(Figure 18B). Samples processed with RIPA buffer had the lowest median % CVs and the method with the lowest median % CV overall was Cap-RIPA-SP3 (19.0%).

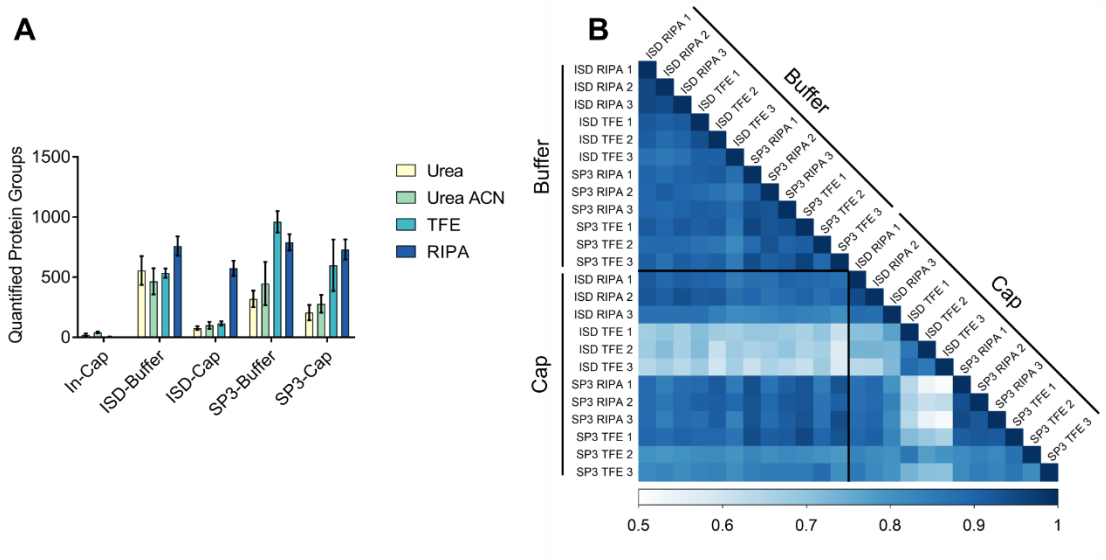

Figure 18 – Quantitative performance of the tested methods

(A) The number of protein groups quantified using MaxQuant's MaxLFQ algorithm (default settings). (B) Pearson correlation matrix of samples where Trifluoroethanol (TFE) or RIPA buffer, scale is truncated for clarity. Cap and Buffer refer to collecting material onto both a dry LCM cap and recovering in the indicated lysis buffer or collecting directly into the indicated lysis buffer. Reproduced from Davis et al. 2019

While using TFE as a protein extraction buffer resulted in the highest numbers of protein and peptide identifications, it is comparatively more difficult to work with than RIPA buffer. This is due to TFE's high vapour pressure causing it to evaporate readily during sample collection and protein extraction, where previously published methods recommend heating at 60 °C for 60 min. Additionally, it does pose a higher safety risk in comparison to the other extraction buffers used here. Considering both these points, the method deemed optimal was to collect samples directly into RIPA buffer and then process them using the SP3 method. This method is among the highest performing when considering identification and quantification quality, alleviates the difficulties of working with TFE, and is amenable to high throughput and automated sample processing.

2.3.3 LCM proteomics reveals differences between human neuronal proteomes

For a proof-of-principle experiment, the method was applied to compare the proteomes of Purkinje and Betz cells, hypothesising that the method would distinguish between the two cell types. Both cell types are easily identifiable neurons that define distinct regions of the human brain – the cerebellar cortex (Purkinje cells) and the primary motor cortex (Betz cells). With this approach, 3551 and 2874 proteins were identified (mean, n = 3) from the Betz cell and Purkinje cell samples, respectively. The approach identified proteins unique to each cell type, with 526 detected only in all Betz cell replicates and 126 detected only in all Purkinje cell replicates. The two cell types are well separated by principal component analysis, with the first principal component driving the separation and contributing 74 % of the variance between the cell types (Figure 19A). Comparing the average LFQ intensity of the proteins within these cells using a two-sample *t* test and plotting the results as a volcano plot reveals a subset of proteins differentially expressed between the two cell types (Figure 19B). 36 % of all proteins detected between the two cell types have a fold-change greater than 2 with a *p*-value of < 0.05.

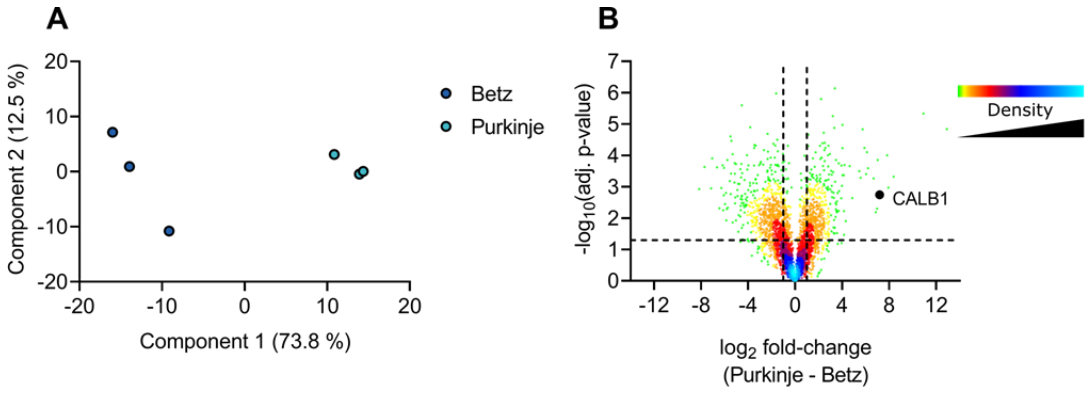

Figure 19 – Proteomic analysis of Purkinje and Betz cells

Analysis of Betz and Purkinje cell proteomes. (A) Principal component analysis of Betz and Purkinje cells based on MaxQuant LFQ values. (B) Volcano plot of the Purkinje cell proteome against the Betz cell proteome. Positive $\log_2$ fold-changes indicate the mean LFQ value is higher in Purkinje cells. Y-axis values are permutation-corrected *p*-values. The horizontal dashed line represents a 5 % FDR threshold. Vertical dashed lines represent ± 2 – fold-change. Colour gradient represents the density of data points. Reproduced from Davis et al. 2019

Some of the proteins with the highest fold-change in the Purkinje cells are known Purkinje cell markers: calbindin 1 (CALB1), a calcium buffering protein essential for normal motor coordination signal output in Purkinje cells [206]; inositol 1,4,5-trisphosphate receptor type 1 (ITPR1), an intracellular calcium channel that is essential for the normal signal output from Purkinje cells [207]; protein kinase C gamma (PRKCG), a calcium-activated kinase that regulates Purkinje cell dendrite growth [208]; and aldolase C (ALDOC), a brain-specific isozyme the aldolase glycolytic enzyme expressed in Purkinje cells [209]. Immunohistochemistry (IHC) was performed for calbindin 1. Purkinje cells stain strongly, as expected, and Betz cells show no staining above the background (Figure 20). There are currently no known Betz cell-specific markers; however, we have validated two proteins upregulated in the Betz cells by IHC: dystonin (DST), a cytoskeletal linker protein essential for the correct organisation of the cytoskeleton of neurons [210]; and ATP synthase F1 subunit beta (ATP5F1B), a catalytic subunit of the mitochondrial ATP synthase complex.

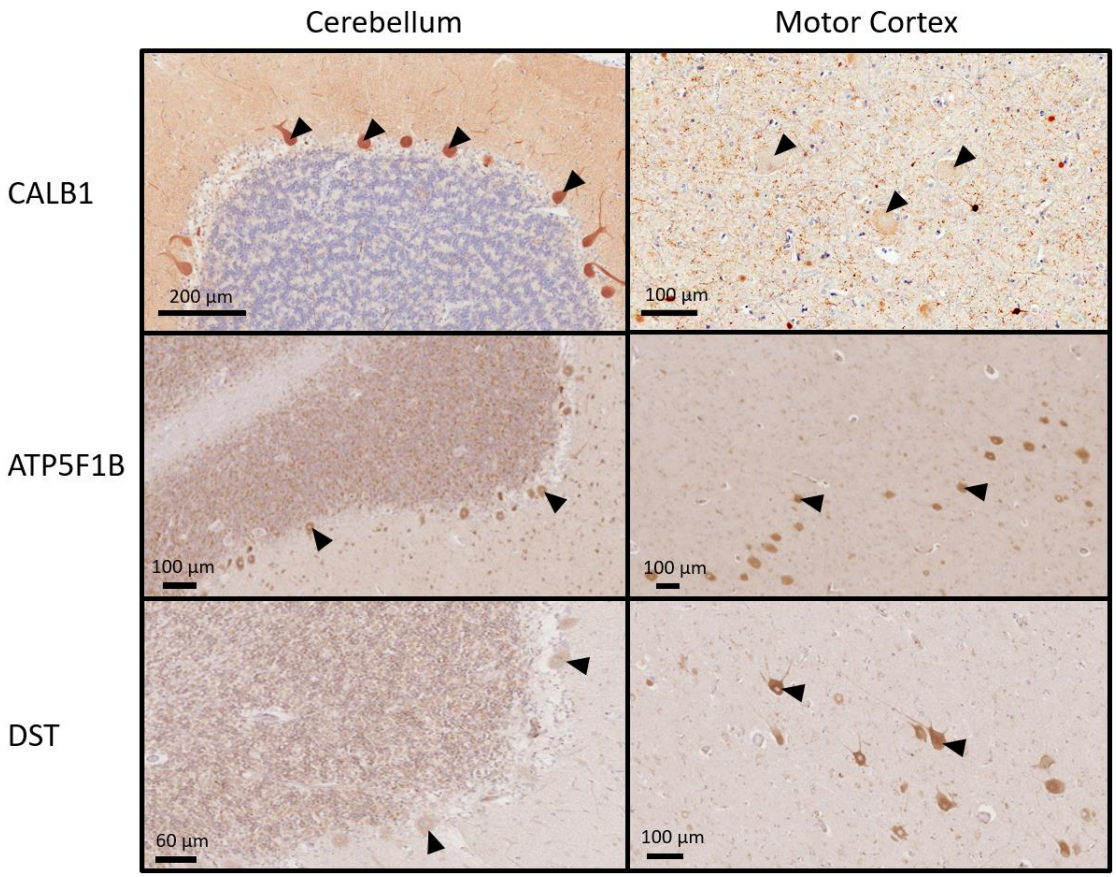

Figure 20 – Immunohistochemistry validation of proteins identified as differentially expressed between Purkinje and Betz cells

Cerebellum and motor cortex stained for calbindin 1 (CALB1), ATP synthase F1 subunit beta (ATP5F1B) and dystonin (DST). Arrowheads indicate examples of Purkinje cells (left, cerebellum) and Betz cells (right, motor cortex). These staining experiments were kindly performed by Connor Scott. Adapted from Davis et al. 2019

In addition to the above validation immunohistochemistry staining, several of the most differentially expressed proteins in Betz and Purkinje cells, labelled in Figure 21, were investigated using publicly available IHC data in the Human Proteome Atlas (HPA)[211].

Table 2 shows the interpretation of IHC staining of human cerebellum tissue samples deposited in the HPA for 11 selected proteins upregulated in Betz cells and 14 selected proteins upregulated in Purkinje cells. Seven of the eleven selected Betz cell proteins showed no detectable staining of the Purkinje cell body, and two proteins showed weak staining (myelin-associated glycoprotein, MAG, and heat shock 70kDa protein 2, HSPA2) of the Purkinje cell body. S100 calcium-binding protein B, S100B, shows intense staining only at the Purkinje cells' cell membrane. All fourteen of the chosen Purkinje cell proteins show positive staining of the Purkinje cell bodies.

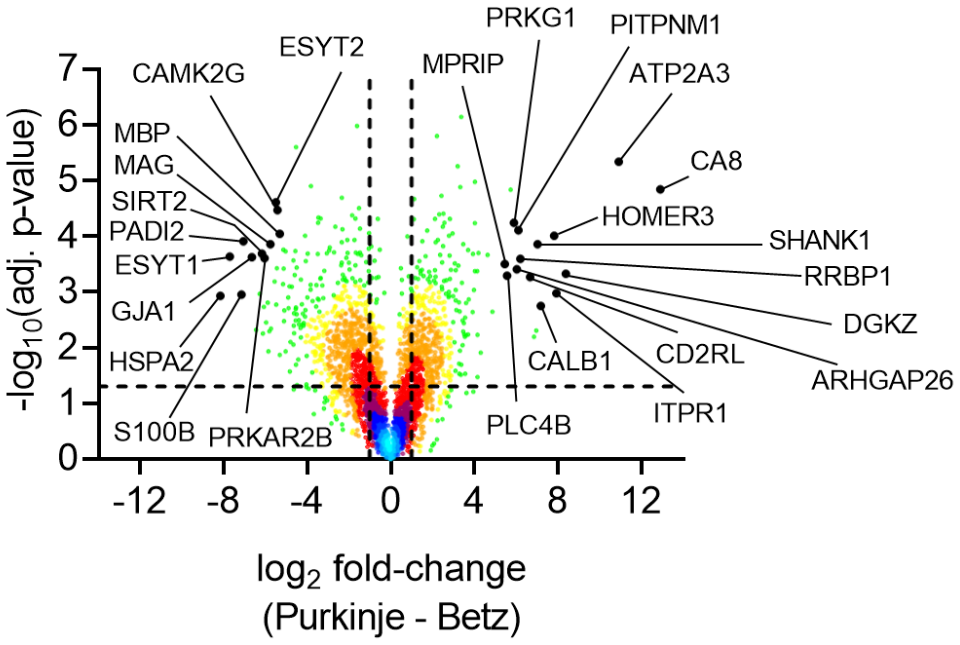

Figure 21 – Differentially expressed proteins correlated to public data.

Volcano plot as in Figure 19 showing proteins differentially expressed between Purkinje and Betz cells. Purkinje cell expression of labelled proteins was investigated in the Human Proteome Atlas cerebellum immunohistochemistry data.

Table 2 – Human Proteome Atlas immunohistochemistry data

Proteins Upregulated in Betz Cells	Human Proteome Atlas Purkinje Cell Staining	Proteins Upregulated in Purkinje Cells	Human Proteome Atlas Purkinje Cell Staining
ESYT2	None	RRBP1	Strong
MBP	None	MPRIP	Moderate
S100B	Strong, membrane only	PITPNM1	Strong
GJA1	None	CALB1	Strong
MAG	Weak	CA8	Strong
PRKAR2B	None	DGKZ	Strong
HSPA2	Weak	PRKG1	Moderate
CAMK2G	Moderate	ITPR1	Strong
SIRT2	None	PLCB4	Weak
ESYT1	None	CDR2L	Strong
PADI2	None	ATP2A3	Strong
		HOMER3	Strong
		ARHGAP26	Strong
		SHANK1	Strong

2.3.4 Comparison to other methods

The optimised method performs excellently in comparison to the existing literature, particularly when considering the metric of proteins per unit of tissue volume isolated and accounting for the LC-MS/MS acquisition time used (1.4 – 2.5 identified proteins/1000 μm^3/gradient hour) (Supplementary Table 1) and is comparable with the application of the nanoPOTS sample processing protocol to rat brain tissue isolated by LCM (2.6 identified proteins/1000 μm^3/gradient hour) [212]. Recently, the nanoPOTS protocol has been adapted to a manual processing-compatible format that still requires the custom nanoPOTS collection & processing slide (µPOTS) [213]. The µPOTS method identified ~2,000 proteins from 40,000 μm^2 of liver tissue using 170-minute gradients (1.8 identified proteins/1000 μm^3/gradient hour). While less sensitive, the advantage of the method presented here over the nanoPOTS/µPOTS workflow is that it does not require the use of custom LCM collection devices, sample preparation robotics, and LC autosamplers. The method presented here performs well in comparison to the nanoPOTS/µPOTS methods while also being more amenable to higher-throughput processing.

2.4 Conclusions

The work presented here shows the development of a sensitive sample preparation method that combines laser capture microdissection with a proteomics workflow. The method can quantitatively detect protein expression differences with cell-type resolution in the spatial context of human brain tissue. Applying the method to cell sections of individual neurons, this optimised method facilitated the detection of protein expression profiles that can distinguish different neuron types from within the same brain tissue. The described methodology lays the groundwork for further investigation of cell physiology/phenotype and function from in-depth proteomic data in the physiological and spatial context of human brain tissue.

3 Spatially resolved proteomics of a human brain tumour

3.1 Introduction

The goal of the work presented in this chapter was to implement the developed workflow for spatially resolved proteomics to a rare paediatric brain tumour – atypical teratoid/rhabdoid tumour (ATRT). Molecular subtyping of ATRT has been performed at the epigenetic and transcriptomic levels[178,182,183]. The proteome of ATRT has not been well explored, with only one large-scale study to date and, until now, no spatially resolved proteomic data have been generated from ATRT tissue[214]. While ATRT does not display stereotyped examples of tumour heterogeneity, like a glioblastoma tumour might, the ATRT case used here displays some tissue heterogeneity (Figure S1).

Most studies that have combined LCM and mass spectrometry-based proteomics have isolated regions of tissue based on existing knowledge such as histological features, immunostaining, gross morphology, and macrostructures[87,138,215]. While this feature-dependent approach is clearly successful for investigating tissue heterogeneity, sampling in an unbiased manner could reveal novel insights into a tissue's spatial protein expression profiles.

While the work in this chapter was being performed, Piehowski *et al.* published a paper where they used LCM to investigate spatial protein expression in the mouse uterus to a depth of 2,000 proteins using the nanoPOTS workflow[139]. They sampled the tissue in an unbiased, gridded manner but used the underlying tissue histology to classify samples before statistical analysis that requires samples to be grouped by a factor. Using this approach, Piehowski *et al.* identified 150 – 175 proteins differentially expressed between three tissue sub-structures. This study is another example of the utility of LCM for investigation tissue heterogeneity but could be extended by combining their unbiased sampling with spatially aware statistical methods which do not require prior knowledge of the underlying tissue.

In this chapter, I show work adapting the developed LCM-Proteomics workflow to a higher-throughput capability. And then, the application of the method to the spatial proteome of a single ATRT tumour sampled using LCM in an unbiased manner to determine spatial expression patterns of proteins within the tumour using spatially aware statistics.

3.2 Materials and Methods

Reagents were purchased from Sigma Aldrich unless otherwise specified. Solvents for LC-MS use were LC-MS grade and purchased from Millipore. Formic acid was purchased from Fisher Scientific.

3.2.1 Tissue retrieval and processing

Post-mortem brain tissue was retrieved by the Oxford Brain Bank; a research ethics committee (REC) approved and HTA regulated research tissue bank (REC reference 15/SC/0639). The retrieved brain was sectioned into 1 cm thick coronal sections starting at the level of the mammillary bodies. Due to its large size, tumour tissue was present within multiple of these coronal slices. The tumour tissue was dissected from the first coronal and second posterior coronal slices (P1 and P2). The tumour from the P2 slice was split into quadrants. Cryosections were taken from all pieces to determine the tissue block with the best morphological and cellular preservation. Cryosections were stained with H&E (see below) and examined by an expert Neuropathologist (Olaf Ansorge). The upper-right quadrant from the P2 coronal slice was selected for use in further experiments.

Relevant tissue blocks of the frozen cerebellum and motor cortex were acclimatised to -20 °C and mounted onto a cryostat block using OCT Compound (Cell Path, ARG1180). Careful consideration was taken to ensure cut sections were not contaminated with OCT. Sections were cut at 10 µm and mounted onto UV irradiated (254 nm, 30 minutes) 1.0 PEN membrane slides (Zeiss) at -18 °C for LCM or Superfrost glass slides for histology. Sections were then air-dried for several minutes and placed onto a Shandon Linistain for automated H&E staining. Sections were fixed in 70 % denatured alcohol, hydrated, stained with Harris' Haematoxylin,

incubated in 0.4 % acid alcohol, placed in Scot's tap water, and stained with Eosin containing 0.25 % acetic acid with regular washing steps in between. Stained sections were then dehydrated in increasing concentrations of denatured alcohol and air-dried without coverslips and stored at -80 °C until processing by laser-capture microdissection. This tissue sectioning and staining was kindly performed by Connor Scott.

3.2.2 Laser capture microdissection

Areas of tissue analysed were annotated and isolated from the prepared slides using a laser-capture microscope equipped with laser pressure catapulting (PALM Microbeam, Zeiss). Cutting and capturing the annotated tissue areas were performed automatically and used the 10x objective lens. The settings in the control software for cutting were Energy: 43, Focus: 55; and for capturing were Energy 20, Focus -15. Samples were collected into 20 µL RIPA buffer (Pierce #89900) in the cap of 200 µL PCR tubes. Collected samples were immediately placed in dry ice. Samples were stored at -80 °C until further use.

3.2.3 Proteomic sample processing

3.2.3.1 Lysis, reduction and alkylation

Samples were thawed, incubated at room temperature for 30 minutes and briefly centrifuged. Caps were rinsed with 20 µL of RIPA buffer (#89900, Pierce) containing 25 units of Benzonase (E1014, Merck) to collect any remaining tissue and briefly centrifuged, followed by incubation at room temperature for 30 minutes to degrade DNA and RNA. Proteins were reduced by adding DTT to 5 mM and incubated at room temperature for 30 minutes, followed by the addition of iodoacetamide to 20 mM and incubation at room temperature for 30 minutes.

3.2.3.2 SP3

Paramagnetic SP3 beads (GE45152105050250 & GE65152105050250, Cytiva) were prepared as described by Hughes *et al.* [22]. Three µL of SP3 beads were mixed with the samples, and acetonitrile added to a final concentration of 70 % (v/v). Samples were mixed with 1000 rpm orbital shaking for 18 minutes, followed by bead immobilisation on a magnet for 2 minutes.

The supernatant was discarded, and beads were washed twice with 70 % (v/v) ethanol in water and once with 100 % acetonitrile without removal from the magnet. Beads were resuspended in 50 mM ammonium bicarbonate containing 25 ng of Trypsin (V5111, Promega) and digested overnight at 37 °C. After digestion, the beads were resuspended by bath sonication. Acetonitrile was added to the samples to 95 % (v/v) and shaken at 1000 rpm for 18 minutes. Beads were immobilised on a magnet for 2 minutes, and the supernatant discarded. Beads were resuspended in 2 % acetonitrile and immobilised on a magnet for 5 minutes. Peptides were transferred to either glass LC-MS vials containing formic acid in water, resulting in a final formic acid concentration of 0.1 %.

3.2.3.3 S-Trap

Samples were brought to 5 % SDS by adding 10 % SDS, 100 mM triethylammonium bicarbonate (TEAB) pH 7.5. Then, 12 % phosphoric acid was added to a final concentration of 1.2 %. Proteins were precipitated by adding 280 µL 90 % methanol, 100 mM TEAB and were transferred to the S-Trap micro cartridges (Protifi). Cartridges were centrifuged at 4,000 *g* until all solution had passed through. Cartridges were washed three times with 90 % methanol, 100 mM TEAB. One µg of trypsin in 50 mM TEAB was added to the top of the S-Trap and was incubated at 47 °C for 1 hour. Peptides were eluted from the S-Trap by centrifugation at 4,000 *g* with 50 mM TEAB, 0.2 % formic acid and then 50 % acetonitrile 0.2 % formic acid. Elutions were pooled and dried to completeness in a vacuum centrifuge and stored at -20 °C until analysis by LC-MS/MS. Peptides were resuspended in 2 % acetonitrile, 0.1 % formic acid immediately before analysis by LC-MS/MS.

3.2.4 LC-MS/MS

Peptides were analysed by LC-MS/MS using two different systems.

Dionex/Lumos:

Peptides were analysed by nano-UPLC-MS/MS using a Dionex Ultimate 3000 coupled to an Orbitrap Fusion Lumos (Thermo Scientific) using a 75 µm x 500 mm C18 EASY-Spray Columns

with 2 µm particles (Thermo Scientific) at a flow rate of 250 nL/min. A 60-minute linear gradient from 2 % buffer B to 35 % buffer B (A: 5 % DMSO, 0.1 % formic acid in water. B: 5 % DMSO, 0.1 % formic acid in acetonitrile). MS1 scans were acquired in the Orbitrap between 400 and 1500 m/z with a resolution of 120,000 and an AGC target of 4×10^5. Precursor ions between charge state 2+ and 7+ and above the intensity threshold of 5×10^3 were selected for HCD fragmentation at a normalised collision energy of 28 %, an AGC target of 4×10^3, a maximum injection time of 80 ms and a dynamic exclusion window of 30 s. MS/MS spectra were acquired in the ion trap using the rapid scan mode.

EvoSep/TimsTOF:

For high-throughput analysis of peptide samples, samples were loaded onto C18 EvoTips before analysis using an EvoSep One (EvoSep) coupled to a timsTOF Pro (Bruker) using a 100 µm x 80 mm C18 column with 3 µm particles (PepSep) and a CaptiveSpray emitter (Bruker). Solvent A was 0.1 % formic acid in water and Solvent B was 0.1 % formic acid in acetonitrile. All centrifugation steps were 700 *g* for 60 seconds unless stated otherwise. For peptide loading onto EvoTips, tips were first washed with 20 µL of Solvent B by centrifugation. Washed tips were conditioned by soaking in 1-propanol until the C18 material appeared pale white. Conditioned tips were equilibrated by centrifugation with 20 µL Solvent A. Samples were then loaded into the tips while the tips were soaking in Solvent A to prevent drying, samples were then bound to the C18 material by centrifugation. Tips were washed by centrifuging with 20 µL Solvent A. Next, 100 µL Solvent A was added to the tips and the tips were centrifuged at 700 *g* for 10 seconds. Samples were then immediately analysed by LC-MS/MS.

The EvoSep One's 100 samples per day method was used, which results in an 11.5-minute gradient at a flow rate of 1.5 µL/min. The TimsTOF Pro was operated in parallel accumulation, serial fragmentation (PASEF) mode. The TIMS accumulation and ramp times were set to 100 ms, and mass spectra were recorded from 100 – 1700 m/z, with a 0.85 – 1.30 Vs/cm^2 ion

mobility range. Precursors were selected for fragmentation from an area of the full TIMS-MS scan that excludes the majority of ions with a charge state of 1+. Those selected precursors were isolated with an ion mobility dependent collision energy, increasing linearly from 27 – 45 eV over the ion mobility range. Three PASEF MS/MS scans were collected per full TIMS-MS scan, giving a duty cycle of 0.53 s. Ions were included in the PASEF MS/MS scan if they met the target intensity threshold of 2000 and were sampled multiple times until a summed target intensity of 10000 was reached. A dynamic exclusion window of 0.015 m/z by 0.015 Vs/cm2 was used, and sampled ions were excluded from reanalysis for 24 seconds.

3.2.5 Mass spectrometry data analysis

3.2.5.1 Protein identification and label-free quantitation

Raw data files were searched against the Uniprot human database (Retrieved 17/01/2017, 92527 sequences) using MaxQuant version 1.6.14.0, allowing for tryptic specificity with up to 2 missed cleavages. Cysteine carbamidomethylation was set as a fixed modification. Methionine oxidation and protein N-terminal acetylation were set as variable modifications and the "match between runs" option was used. All other settings were left as default. Label-free quantification was performed using the MaxLFQ algorithm within MaxQuant. Protein and peptide false discovery rate (FDR) levels were set to 1 %.

3.2.6 Spatial data analysis

3.2.6.1 Feature-independent analysis

The feature-independent spatial analysis uses functions within the spdep and raster R packages[141,216,217]. MaxQuant's protein level output files ('proteingroups.txt') are read into an R data frame and filtered to remove reverse hits, 'Only identified by site' hits and potential contaminants. The 'LFQ intensity' columns are $\log_2$ transformed and then normalised by median subtraction. Protein groups that do not meet a cut-off of having at least 9 pixels with normalised LFQ values are not taken forward for further analysis. The following steps occur independently for each protein group. Normalised LFQ intensities are then coerced into a matrix that reflects the raster pattern of sample acquisition. The quantification matrix is

converted into a raster object and then to a polygon object using the Raster package. From this polygon object, a neighbour list is built for each cell (pixel) of the raster using the 'Queen's Case' where cells are considered neighbours if they share an edge or a vertex. The neighbour list is then supplemented with a spatial weights matrix using a binary coding scheme where neighbours are given a weighing of '1' and non-neighbours a weighting of '0' in the spatial weights matrix. The raster object and the weighted neighbour list are then used as inputs to a permutation test for the Moran's *I* statistic, calculated using 999 random spatial permutations of the raster object to calculate pseudo-*p*-values. Moran's *I* statistics and the associated *p*-values are collected for every protein group. The *p*-values are then corrected for multiple testing using the Benjamini-Hochberg FDR method.

For local Moran's *I* testing, $\log_2$ transformed, and median normalised protein LFQ values were processed to the raster and weighted neighbour list objects as above. These objects are then used as inputs to the localmoran function of spdep to calculate the local Moran's *I* statistic and associated Benjamini-Hochberg FDR-adjusted *p*-values for each pixel independently for each protein group. The protein intensity variable was then plotted against a spatially-lagged version of that variable to generate a Moran scatterplot. Samples were classified as hotspots, coldspots or spatial outliers based on the quadrant they are located in if they were significant in the local Moran's *I* test.

3.2.6.2 Feature-dependent analysis

Feature-dependent analysis was carried out in Perseus (version 1.6.14.0). MaxQuant's protein level output files ('proteingroups.txt') and filtered to remove reverse hits, 'Only identified by site' hits and potential contaminants. The 'LFQ intensity' columns are $\log_2$ transformed and then normalised by median subtraction. Samples were classified based on the Neuropathologist's expert annotation of the tissue section. Two-sample *t*-tests were performed using a permutation-based multiple testing correction FDR of 5 %. A $\log_2$ fold-change threshold of ± 1 (fold change ± 2) was used to filter the significant hits.

3.2.7 Pathway analysis

Protein groups called as significant by the feature-dependent analysis were taken forward for pathway analysis using clusterProfiler using a background list of all proteins detected within the respective experiment[218,219]. The list of significant protein groups was split into two based on the direction of the $\log_2$ fold-change. Protein identifiers were converted to Entrez gene identifiers[220]. The two lists were tested for enrichment of Reactome pathway terms and Gene Ontology Biological Process terms using a Fisher's exact test with a 5 % FDR threshold.

3.2.8 Cell counting

H&E-stained slides were imaged using Aperio Scanscope Turbo AT (Leica) with a 20x/0.75 NA Plan Apo objective lens coupled with an x2 optical magnification lens to achieve a total magnification of 40x and a resolution of 0.25 µm/pixel. Digitalisation was achieved by scanning strips of tissue, which were then stitched together until the whole section was completed. Scanned images were loaded into QuPath (version 0.2.0m11)[221]. The cell detection tool was used to detect and count nuclei within annotated regions using the default parameters using the Haematoxylin optical density (OD) as the detection image.

3.3 Results and Discussion

The spatial proteomic workflow performed uses LCM to capture regions of tissue directly into lysis buffer in a contactless manner before proteomic sample preparation and digestion using the SP3 method and analysed by LC-MS/MS. Protein intensities can be visualised within the spatial context of the sampled tissue, and proteins exhibiting spatial expression patterns determined (Figure 22).

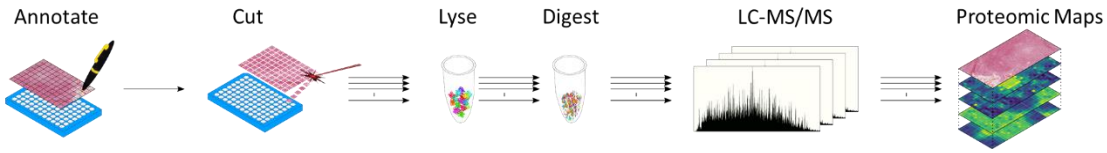

Figure 22 – Spatially resolved proteomic workflow

A grid of equally sized, square regions (raster pattern) is annotated onto the tissue in the LCM software. These regions are then isolated into sample processing tubes or plates containing lysis buffer and processed by a modified SP3 method before tryptic digestion. The resulting peptides are analysed by LC-MS/MS. Raw data are searched, the identified proteins quantified and then mapped back to the original raster positions from the annotation.

During the time that the work presented in the previous chapter was being performed, a proteomic sample preparation technique called S-Trap became popular within the proteomics community for its ease of use, compatibility with detergent, and performance[21,222]. Before proceeding with experiments on the ATRT tumour, this new method was tested against the

optimised method from the previous chapter. 100,000 μm^2 of 10 µm thick ATRT tissue was collected into RIPA buffer in triplicate and processed by either SP3 or S-Trap (Figure 23). On average, 3,117 and 2,372 proteins were identified in the samples prepared by SP3 and S-trap, respectively (mean, n = 3). The same pattern of higher identifications by SP3 was repeated on the peptide level, with 17,995 and 10,860 peptides identified by SP3 and S-trap, respectively (mean, n = 3). Analysis using a two-way ANOVA results in a significant difference between the two methods ($p < 0.001$). The effect of freeze-thaw cycles on the tissue was checked in the same experiment due to the possibility of sampling tissue from the same section over multiple days or to revist a section later. No significant difference in the number of protein identifications between zero freeze-thaw cycles and one freeze-thaw cycle after H&E staining (which itself gives causes a freeze-thaw cycle) was found for either samples processed by SP3 or S-Trap (two-way ANOVA, $p = 0.835$) (Supplementary Table 2).

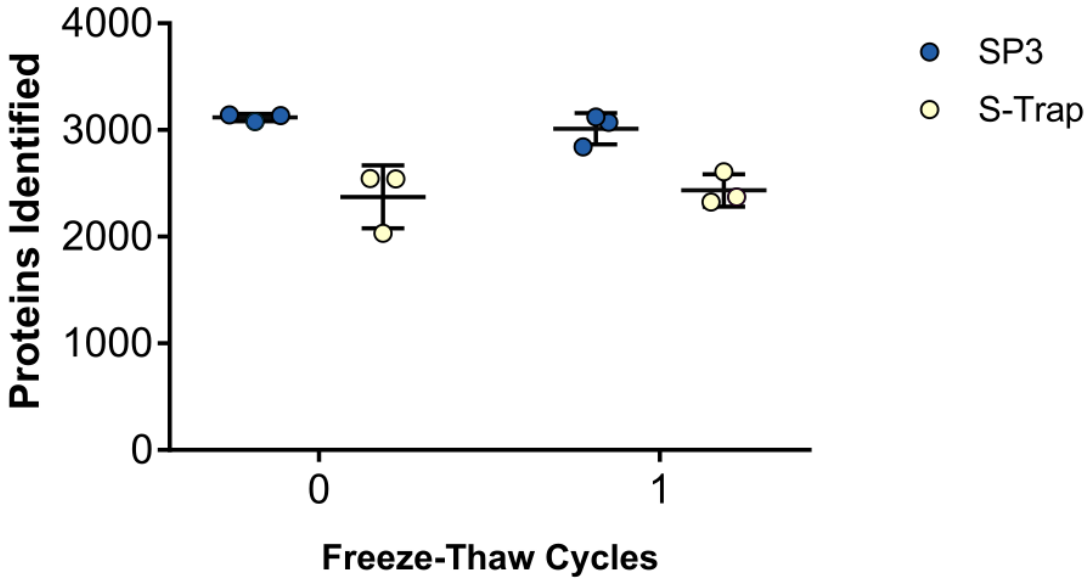

Figure 23 – Protein identifications in ATRT processed with SP3 and S-trap

The number of proteins identified from 100,000 µm^2 ATRT tissue digested with either the optimised SP3-based protocol or S-Trap and after 0 or 1 additional freeze-thaw cycles. The horizontal bar represents the mean, and error bars represent the standard deviation. From a two-way ANOVA test, SP3 identifies significantly more proteins than S-Trap ($p < 0.001$), and the difference between 0 and 1 freeze-thaw cycles is not significant ($p = 0.835$).

3.3.1 Preliminary spatial proteomics sensitivity analysis

Biological structures exist over many orders of magnitude, and most are accessible via laser capture microdissection, from single cells (≈ 25 µm^2) upwards. The next step was to characterise how the number of protein identifications changes depending on how much tissue is collected. A titration experiment was performed where $10^{0.5}$ µm^2 (3 µm^2) of tissue, representing virtually no material, to 10^6 µm^2 (1,000,000 µm^2 [1 mm^2]) were isolated with semi-logarithmically increasing steps of area (Figure 24). These samples were digested and analysed using a Dionex U3000 LC system connected to an Orbitrap Fusion Lumos with a 60-minute linear gradient where peptides elute from the LC column for analysis, capable of analysing ⌣13 samples per day (referred to as Dionex/Lumos). The range of protein identifications per sample was from 282 – 3,480. Until an area of $10^{3.5}$ µm^2 (3,162 µm^2), the

number of identified proteins was around 325, representing a background level of proteins. From $10^{3.5}$ µm^2, the number of proteins identified increases with an increase in the area of tissue collected. A slight tail-off begins at $10^{5.5}$ µm^2 (316,000 µm^2) and 10^6 µm^2 (1,000,000 µm^2 [1 mm^2]), likely caused by approaching the maximal amount of MS/MS spectra that can be acquired in the fixed gradient time of 60 minutes. This defines a lower limit on the range of sampling the ATRT tissue: $10^{3.5}$ µm^2 (3,162 µm^2)and an approximate upper-threshold were increasing area collected starts to give diminishing returns: $10^{5.5}$ µm^2 (316,000 µm^2). These areas correspond to squares with a side length of 56 µm and 560 µm, respectively. These numbers of protein identifications again compare favourably to other work. Zhu *et al.* [212] performed LCM of discs of rat brain 200 µm in diameter and 12 µm thick, (approximately equivalent to 37,000 µm^2 of 10 µm thick tissue used here), and identified 1,827 proteins – the $10^{4.5}$ µm^2 (31,000 µm^2) sample here identified 1,727 proteins.

Naturally, as the sample area decreases, the number of samples required to cover a fixed area will increase, so the amount of LC-MS/MS analysis time & cost will increase. Future work mapping spatial proteomes could require hundreds to thousands of samples, the same titration experiment approach was performed and analysed on a modern high-throughput LC-MS/MS system: a timsTOF-pro coupled to an EvoSep One running 12-minute gradients, capable of analysing ∽100 samples per day (referred to as EvoSep/TimsTOF) [33,45]. The number of protein identifications was lower in general, ranging from 20 – 1,419. With this LS-MS/MS system, the signal only rose above the (albeit lower) noise at $10^{4.5}$ µm^2 (31,000 µm^2) with 636 proteins identified, then immediately reaching a plateau at 10^5 µm^2 (100,000 µm^2) with 1,345 proteins identified. These two systems, Dionex/Lumos & EvoSep/Tims, represent the standard configurations of the two highest-end mass spectrometers within our laboratory. While literature shows that sensitivity can be increased with a lower LC flow rate and column diameter [192], the aims of the spatial proteomic work require robust analysis of large numbers

of samples, a feature already proven on these setups through their routine use on a wide range of samples as part of the work performed in our laboratory.

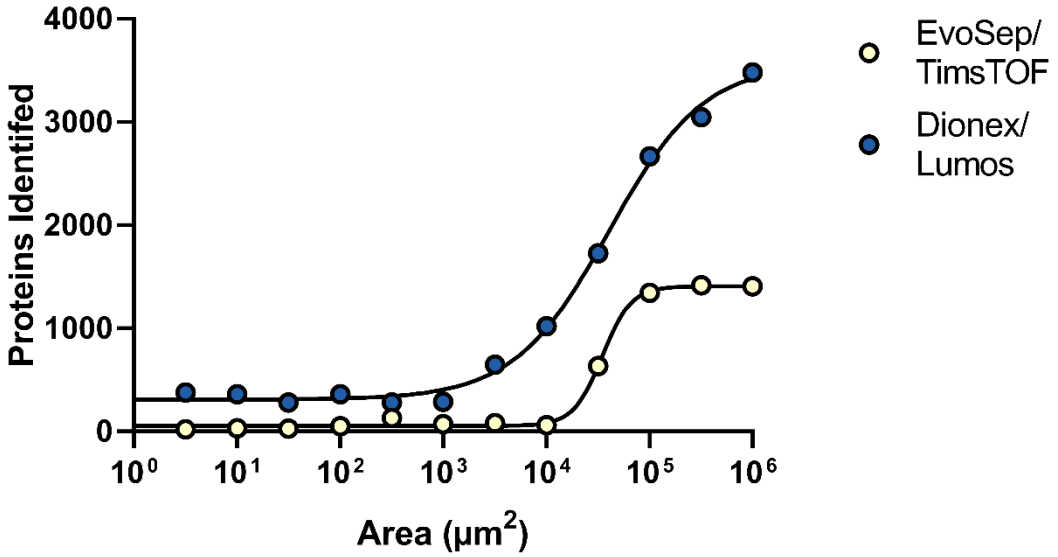

Figure 24 – Protein identifications on two LC-MS/MS systems from a range of ATRT areas.

The number of proteins identified from a titration of ATRT area on two LC-MS/MS systems. Dionex/Lumos refers to a throughput of ∽13 samples/day on an Orbitrap Fusion Lumos coupled to a Dionex U3000 running at 250 nL/min. EvoSep/TimsTOF refers to ∽100 samples/day on a timsTOF-pro coupled to an EvoSep One running at 2 µL/min.

3.3.2 Pilot ATRT spatial proteomics experiment

After characterising the 'response curve' for the ATRT tissue, a pilot spatially resolved proteomics experiment was performed (Figure 25). An area of approximately 1.1 cm^2 (∽4.1 mm x ∽2.7 mm) was subdivided into 96 (12 x 8) square 'pixels' with a side length of 346 µm and an area of 120,000 µm^2 from a 10 µm thick tissue slice. The region of tissue sampled was annotated by a neuropathologist (O.A) who identified two major areas; an area of solid tumour containing neoplastic cells and a region termed 'brain/tumour interface', which contained a mixture of normal and neoplastic cells and three prominent blood vessels. Each pixel was isolated and collected by LCM and processed using the optimised LCM-SP3-Proteomics protocol. This size was chosen to allow for the resulting peptides to be split in half and analysed on the Dionex/Lumos and EvoSep/TimsTOF systems. The equivalent area

injected on each system was 60,000 µm^2 (equivalent side length of 245 µm) as this area is at the inflexion point of the EvoSep/TimsTOF system.

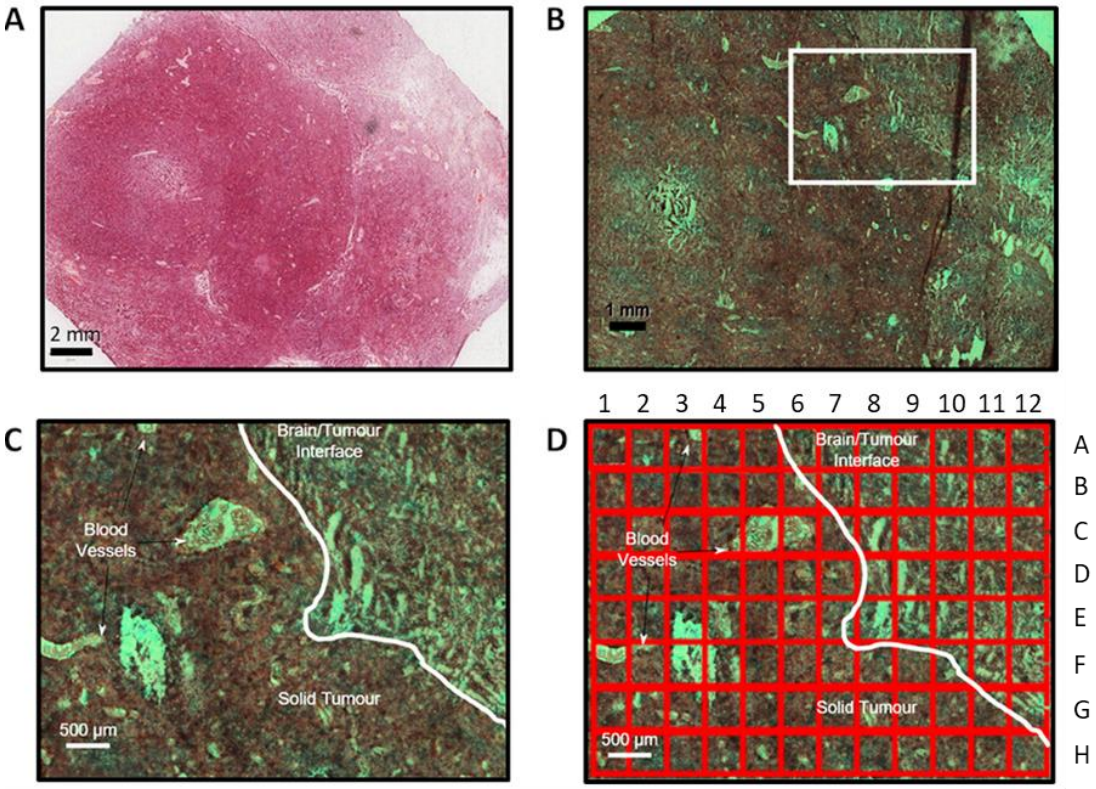

Figure 25 – ATRT Histology and region used for raster sampling

(A) Whole-slide image of H&E stained ATRT section. (B) Composite image of H&E stained ATRT section stitched from multiple fields of view acquired directly on the LCM instrument. (C) Zoomed region highlighted in the white box of B with the expert Neuropathologist's annotations. (D) As C, with raster sampling LCM annotations overlaid. Each pixel of the grid was isolated, digested and analysed by LC-MS/MS. H&E staining was kindly performed by Connor Scott

In total, the Dionex/Lumos system identified 49,091 peptides corresponding to 3,950 proteins with a wide range of proteins identified per sample: 693 - 29857 peptides corresponding to 222 – 3,478 proteins. The total ion current (TIC) of eight of the samples (C8 – D4) showed evidence of contamination with polymeric polyethylene glycol (PEG), evident by clusters of peaks separated by +44.05 Da, a mass shift equal to the PEG monomer, and so were excluded from further analysis (Figure S2). Because of the range of proteins identified per sample, only 38 proteins were identified in every sample. From these sets of identified proteins, 3,679 were

quantified in at least one sample, and 20 proteins were quantified in all samples. Figure 26A shows the number and distribution of proteins quantified per sample. On plotting the distribution of proteins quantified per sample, three populations are visible. The 'High' cluster (58 samples) showed on average 2,485 quantified proteins, the 'Medium' cluster (7 samples) showed on average 1,537 quantified proteins, and the 'Low' cluster (23 samples) showed on average 441 quantified proteins. As the experiment was conducted with spatial resolution, these identified clusters can be easily mapped back to the tissue isolated to check for any correlation with features present (Figure 26B). The samples with lower numbers of quantified proteins do not form strong patterns or correlate with the underlying tissue pathology from the H&E stain (Chi-square test, $p = 0.80$).

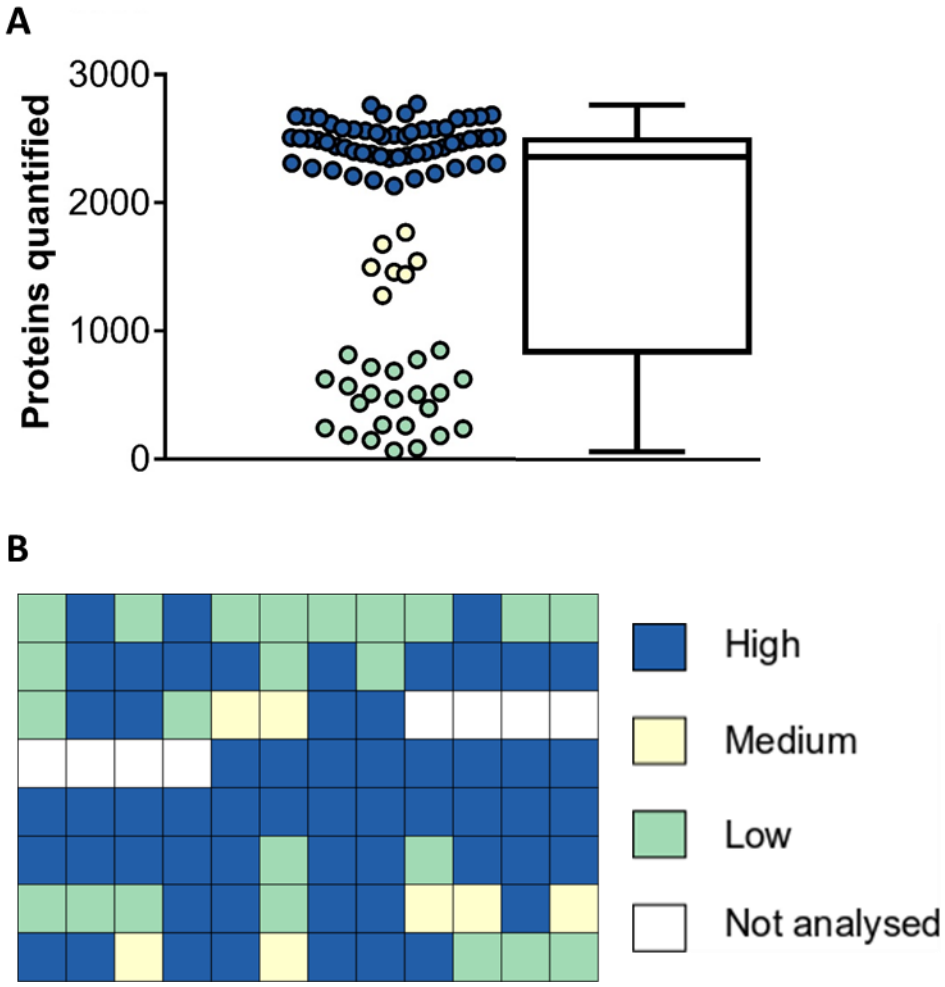

Figure 26 – Distribution of proteins quantified per samples in the ATRT pilot experiment.

(A) Distribution of quantified proteins per sample. Samples are coloured based on which of the three visible clusters they fall in: High (58 samples), Medium (7 samples) and low (23 samples). Data points are jittered on the x-axis to avoid overplotting. (B) Cluster membership determined in (A) mapped back to the raster sample pattern. Eight samples were not included in the analysis due to PEG contamination.

3.3.2.1 Feature-dependent analysis

As the dominant tissue type of each sample can, in this case, be determined from a histology image, statistical comparisons such as t-tests and ANOVA can determine proteins with statistically significant differential expression. Figure S3 shows a clear separation by PCA between the samples with lower numbers of proteins quantified ('Low' cluster) and the rest of the samples. After removing the 'Low' cluster of samples, clear separation between the solid tumour pixels, brain/tumour interface pixels and the two pixels covering the large central

blood vessel is visible after PCA (Figure 27A). A two-sample *t*-test identifies 135 differentially expressed proteins: 91 proteins as significantly increased in the solid tumour and 44 proteins significantly decreased in the solid tumour when compared to the brain/tumour interface with an adjusted *p*-value < 0.05 and a $\log_2$ fold-change threshold of 1 (Figure 27B). Samples that have the annotation line passing through them were excluded from the analysis. Two examples of significantly differentially expressed proteins from this test are glycogen phosphorylase, liver form ($\log_2$ fold-change = 4.5, $q < 0.001$) and peripherin ($\log_2$ fold-change = -7.0, $q < 0.001$). Neither of haemoglobin or histone H4 are significantly differentially expressed between the solid tumour and brain/tumour interface ($\log_2$ fold-change = -0.70, $q = 0.14$; $\log_2$ fold-change = -0.18, $q = 0.20$).

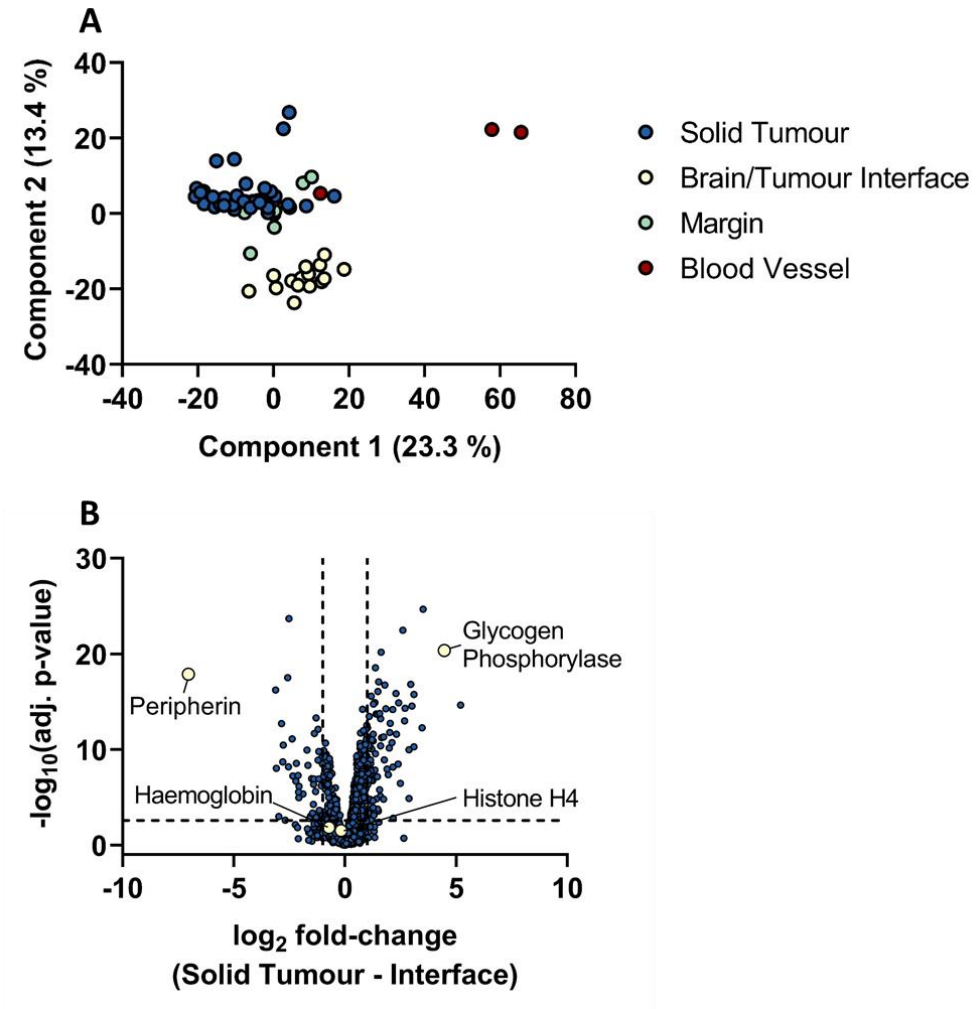

Figure 27 – Feature-dependent analysis of ATRT spatial proteomic data

Analysis of ATRT spatial proteomic data after filtering samples with a low number of proteins quantified. (A) Principal component analysis of ATRT tissue pixels based on MaxQuant LFQ values. Legend represents the classification of the pixel-based on annotated histology image. Margin pixels are pixels that contain areas of both solid tumour and brain/tumour interface. (B) Volcano plot of ATRT tissue pixels classified wholly as solid tumour or brain/tumour interface. Positive $\log_2$ fold-changes indicate the mean LFQ value is higher in solid tumour. The horizontal dashed line represents a 5 % FDR threshold. The vertical dashed lines represent ± 2-fold change.

Peripherin (PRPH) is a type III intermediate filament (IF) protein expressed mainly in peripheral nervous system neurons and central nervous system neurons with peripheral axons[223,224]. The specific function of peripherin is not yet clear. Peripherin is expressed highly during development and after nerve injury, potentially playing a part in axon guidance and regeneration, and yet an alternatively spliced isoform is upregulated in Amyotrophic Lateral

Sclerosis (ALS), a neurodegenerative disease[225–227]. As peripherin is expressed mainly in neurons that are differentiating and undergoing axon growth, the expression pattern in Figure 28 is consistent with the histopathological classification of tumour growth into surrounding normal tissue[228].

A lack of differential expression for histone H4 (and other histone protein variants) is consistent with the low number of genetic aberrations in ATRT in general. ATRT does not show high levels of aneuploidy, and as the mass of histone proteins per cell scales is approximately equal to the mass of DNA per cell, it would be expected that as that mass of DNA stays consistent between ATRT cells and normal cells, so would the mass of histone protein[183,229–231].

Glycogen phosphorylase is the enzyme responsible for the release of glucose from glycogen before entry into glycolysis. This is via the release of glucose-1-phosphate and exists as three isoforms: a liver form (PYGL), a brain form (PYGB) and a muscle form (PYGM). Glycogen catabolism is a vital process in normal brain tissue involved in learning, memory formation, K^+ and glutamate uptake, and glutamine neurotransmitter synthesis[232–235]. Altered brain metabolism of glycogen occurs in a range of diseases such as Alzheimer's disease, epilepsy and diabetes[236]. PYGB is found mainly in astrocytes in the healthy brain but is also expressed at low levels, concomitant with low glycogen levels, in neurons[237,238]. This allows the neurons to tolerate periods of hypoxia and hypoglycemia. Glycogen phosphorylase enzymes show an increased expression in response to hypoxia and nutrient deficiency, and the hypoxia-inducible factors (HIFs) are important transcription factors regulating this response. In cancer, glycogen phosphorylase expression is associated with malignant phenotypes and a poorer prognosis and is a crucial driver of cancer cell growth, survival and hypoxia resistance and is upregulated in metastatic cells [239].

In this data, both the brain and the liver isoforms of glycogen phosphorylase are detected, but only the liver form is differentially expressed across the tissue. Although the isoforms catalyse

the same reaction, they are under different regulatory control mechanisms. For instance, PYGB is allosterically activated by AMP while PYGL activity is insensitive to AMP regulation and is driven by phosphorylation[240]. Mathieu *et al.* show that the addition of H_2O_2 inhibits the activity of PYGB but not PYGL[241]. This is through a redox-dependent switch that controls the allosteric activation of PYGB by AMP, the oxidising conditions cause the formation of a disulphide bridge within the AMP-binding site[241]. Furthermore, PYGL has been shown to display a similar expression pattern to known genes regulated by hypoxia and hypoxia drives the production of oxidising reactive oxygen species[242,243]. Here, this upregulation of PYGL within the area of solid tumour could be in response to the tumour cells' energy demands not being met within a hypoxic, ROS-generating (and PYGB-inhibiting) environment. The total intensity of PYGL across the region is four times higher than that of PYGB, potentially indicating that PYGL is the dominant isoform. However, the proteomic analysis used here is best suited for characterising the abundance of proteins across different samples, rather than between different proteins in the same samples, so it cannot confidently determine which isoform is truly more abundant.

3.3.2.1.1 Proteomic maps

As the spatial location of each sample relative to others is preserved with the rastered sampling technique, the protein LFQ abundances can be plotted in the relative spatial locations to visualise the distribution of the proteins quantified in the experiment as a map of protein expression. Figure 28 shows these protein maps for PRPH, PYGL, histone H4 and haemoglobin in the context of the pathologist's annotation. The expression of PRPH is generally higher in the brain/tumour interface region, and the expression of PYGL is generally lower. The high expression of PYGL extends past the annotated margin slightly, still showing higher expression in pixels that are annotated as mainly consisting of brain/tumour interface. Unsurprisingly, haemoglobin abundance is highest in the two pixels covering the sizeable

central blood vessel. Histone H4 has a more consistent expression across the region. The only visible pattern mirrors the distribution of proteins quantified per sample as in Figure 26, i.e. pixels with a low or medium number of quantified proteins display a lower intensity histone H4. Potentially, this difference could be due to a difference in cell number between the different pixels, so the distribution of nuclei in an adjacent H&E-stained slice was investigated and was found to be consistent between elements of a grid the same size as used here, with a mean number of detected nuclei of 710 per 350 µm x 350 µm grid.

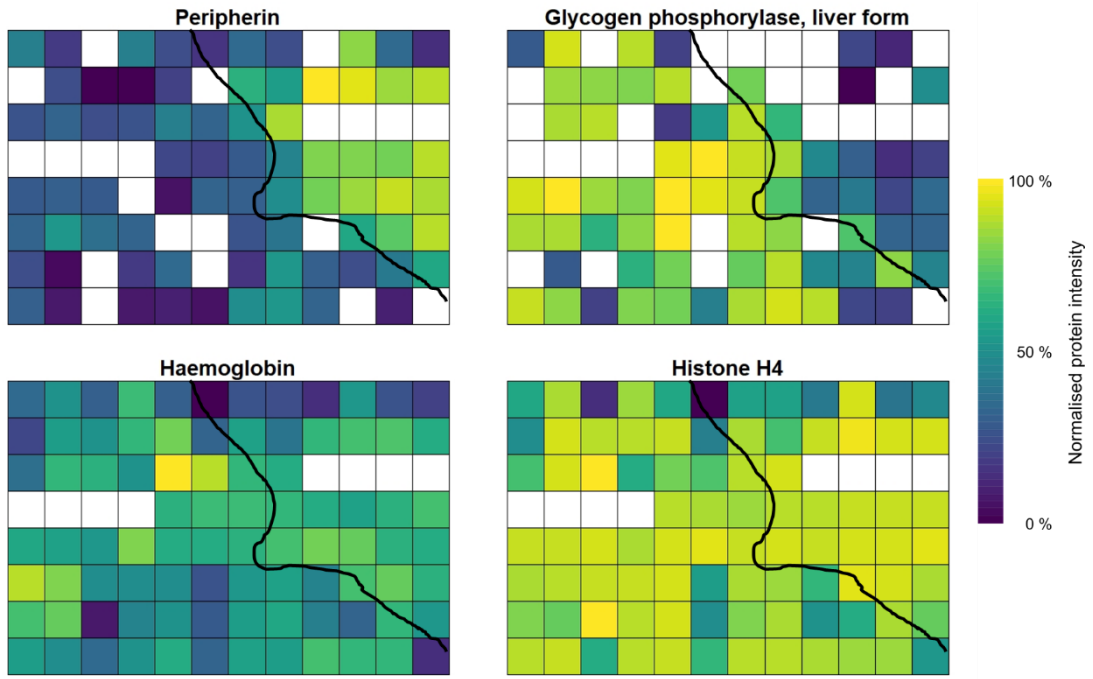

Figure 28 – Spatial proteomic maps of four example proteins

Normalised protein intensity of four example proteins mapped to the spatial positions of tissue isolated using LCM, each pixel represents one sample. Black line denotes the position of the margin between the solid tumour (left of line) and the brain/tumour interface (right of line) as determined by the expert Neuropathologist. Normalised protein intensities are scaled separately for each protein. White = not quantified.

3.3.2.1.2 Feature-dependent pathway analysis

To investigate functional and pathway-level differences between the two annotated regions of solid tumour and brain/tumour interface, the 135 differentially expressed proteins were

annotated with Gene Ontology (GO) biological process and Reactome pathway terms and tested for over-representation as compared to all proteins identified in the experiment using a Fisher's exact test (Figure 29, a more global overview is displayed in Figure S4-7). This analysis reveals an enrichment of proteins involved the organisation of extracellular matrix (driven by enrichment of collagens and laminins), cytoskeletal organisation (driven by enrichment of three tubulin isotypes: tubulin beta 3 class III; tubulin beta 6 class V; tubulin alpha 1c), Rho GTPase signalling via IQGAP1, and neuronal development (mainly consisting of proteins also involved in Rho GTPase signalling, cell motility and cytoskeletal proteins). These functional enrichments within the region of solid tumour are generally consistent with observations that knockdown of SMARCB1 in 293T and MCF7 cells results in a loss of cell adhesion and increased cell migration that was associated with an increase in RhoA activity, and that re-expression of SMARCB1 in malignant rhabdoid tumour (MRT) cell lines reduced cell migration [244].

The functional enrichments within the brain/tumour interface region include cell adhesion via the neuronal cell adhesion molecule L1 (L1CAM), neuronal development via a range of proteins involved in neurogenesis, and transport of amino acids. Additionally, there is enrichment related to fibrin-clot formation driven by prolylcarboxypeptidase (PRPC), glia-derived nexin (SERPINE2) and protein S (PROS). Further investigation of the function of these clotting-related proteins shows that, while involved in fibrin clot formation processes, they are also involved in neuronal differentiation (SERPINE2), neuropeptide regulation (PRCP) and blood-brain barrier integrity (PROS)[245–249].

Taken together, these functional enrichments mirror known features of ATRT and cancer hallmarks such as metabolic reprogramming, altered ECM organisation and cell adhesion in the solid tumour, along with a more neuronal phenotype in the brain/tumour interface region, reflecting the presence of more normal brain tissue.

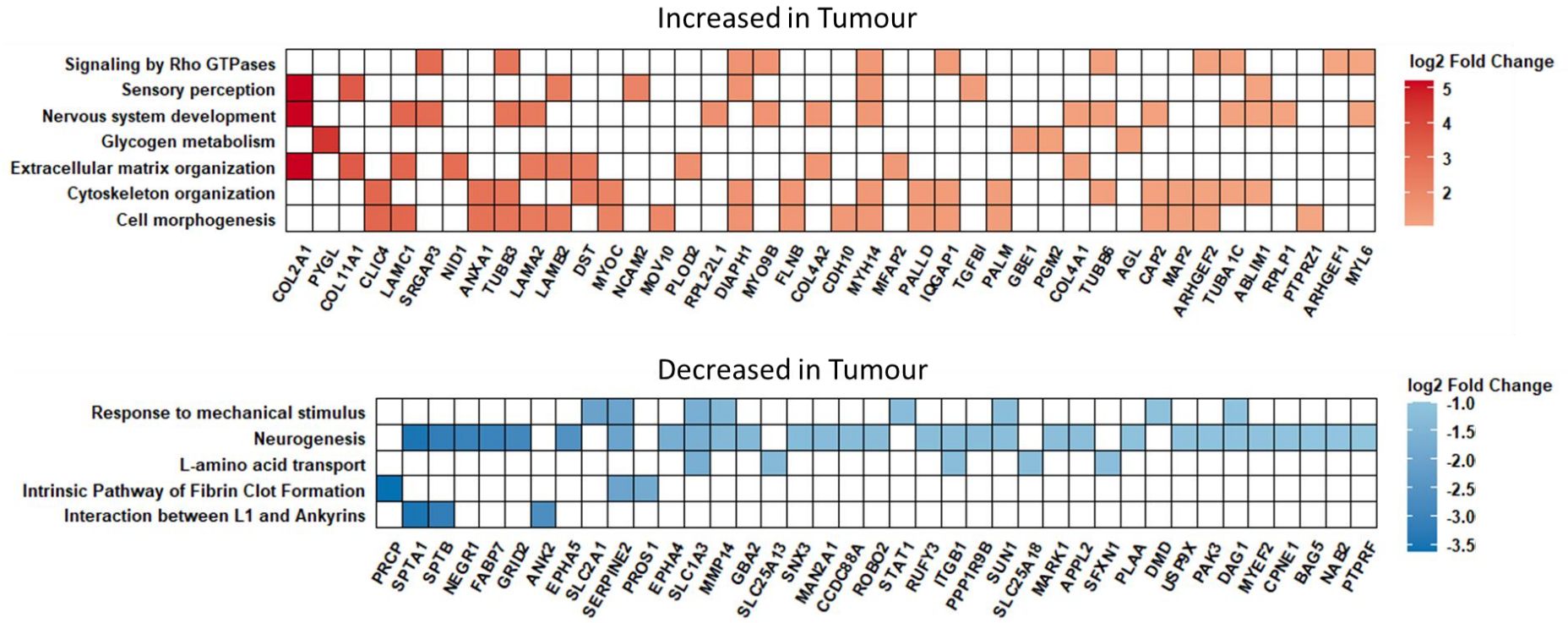

Figure 29 – Functional analysis of feature-dependent ATRT proteomic data

Functional enrichment of Gene Ontology Biological Process and Reactome Pathway terms within differentially expressed proteins between the region of solid tumour and brain/tumour interface. Gene set membership is indicated by colouring the corresponding cell with that gene's log2 fold change.

3.3.2.2 Feature-independent analysis

This approach of defining samples into groups before testing for differentially expressed proteins requires prior knowledge, which may not be available, so a methodology that does not require this definition is a desirable one. Spatially-aware statistical methods are one way that this approach can be extended not to require any prior knowledge other than the relative locations of the sampled areas. In addition, this class of statistical methods can be applied to regions comprised of arbitrary polygons, not necessarily the gridded structure used here. To determine whether a spatial statistics method would be helpful in the investigation of spatially resolved proteomic data, the Moran's *I* test was applied to the proteins quantified in the data from raster sampling. Moran's *I* ranges from -1 to +1, where -1 indicates dispersed data, 0 indicates complete spatial randomness, and +1 indicates perfect clustering of high and low values. The Moran's *I* statistic is a measure of spatial autocorrelation that tries to answer the question: does the spatial distribution of a variable correlate with itself in space more than would be expected by chance? If the answer to that question is yes, then the assumption is that there is some causal reason why the variable is not randomly distributed in space.

Across the entire experiment, 3,679 proteins were quantified in at least one sample. However, for spatial analysis, it is required that the variable of interest is detected in pixels that neighbour each other. Therefore the 3,061 proteins quantified in at least nine samples were taken forward for analysis. To calculate p-values for the Moran's *I* test, the observed value for Moran's *I* is compared against a distribution of Moran's *I* values calculated from random permutations (in space) of the observed protein LFQ values. The p-value represents the rank of Moran's *I* of the non-permuted dataset in comparison to permuted data. As 999 permutations were calculated, a p-value of 0.001 for a given protein represents that the protein's Moran's *I* value is larger than that of all of the permuted values. Conversely, a p-value of 0.5 indicates that 50 % of permuted Moran's *I* values are greater than the value calculated for the actual data. Of the proteins tested, 273 proteins have a p-value = 0.001, 512

proteins have a p-value ≤ 0.01 and 957 proteins have a p-value ≤ 0.05. After applying a Benjamini-Hochberg multiple testing correction, 451 proteins have an adjusted p-value ≤ 0.05, indicating evidence for spatial expression patterns. Returning to the four previous examples of peripherin, PYGL, haemoglobin and histone H4, these proteins have Moran's *I* values of 0.58, 0.41, 0.19 & 0.09 and adjusted p-values of 0.011, 0.011,0.019 & 0.16, respectively. Promisingly, these statistics represent what was seen in the feature-dependent approach (significance for peripherin and PYGL, no significance for histone H4) and what is visible in the protein map for haemoglobin, where the prominent blood vessel is easily visible.

In order to compare the two approaches of feature-dependent and feature-independent, the two sets of proteins called as significant by the methods were compared. Ninety-four proteins have both a significant Moran's *I* (q ≤ 0.05) and a significant *t*-test difference ≥ 1 (& q ≤ 0.05), 357 proteins are called as significant only by the feature-independent analysis, and 46 proteins are called as significant only by the feature-dependent analysis. A more stringent approach to the feature-independent analysis would be to only call proteins with $p = 0.001$ as significant, as none of these protein's Moran's *I* was exceeded with a permuted value; using this criterion, 76 proteins were called as significant by both approaches (Figure 30). Supplementary Table 3, Supplementary Table 4, and Supplementary Table 5 show the top 20 proteins significant sets of proteins in Figure 30B and their corresponding protein maps are shown in Figure S8, Figure S9, and Figure S10, respectively.

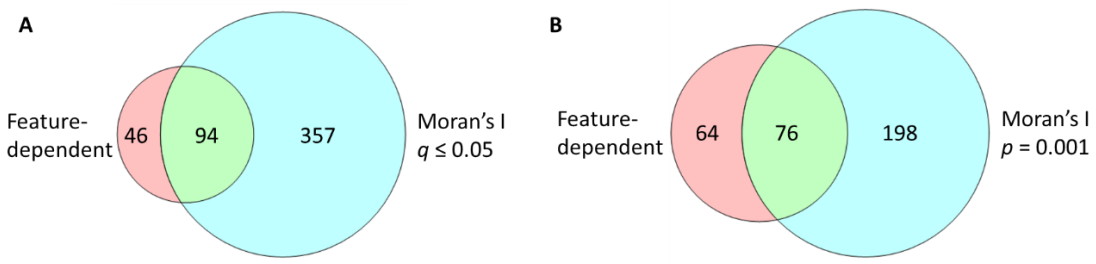

Figure 30 – Overlap of significant proteins between feature-dependent and feature-independent approaches

(A) Overlap of proteins called as significant after *t*-tests between pixels of solid tumour and of brain/tumour interface as determined by a neuropathologist and the annotation-independent approach of spatial autocorrelation analysis using Moran's *I* tests with an FDR ≤ 5 %. (B) Overlap of proteins called as significant after *t*-tests between pixels of solid tumour and of brain/tumour interface as determined by a neuropathologist and the annotation-independent approach of spatial autocorrelation analysis using Moran's *I* tests with a p-value of 0.001 (the smallest p-value obtainable with the parameters used)

To further investigate the feature-independent analysis's behaviour compared to the feature-dependent analysis, the distribution of Moran's *I* values was compared between proteins called as significant by either of the methods. The median Moran's *I* for proteins not called as significantly different by either method was 0.050. For proteins that were called as significant by the feature-dependent approach, the median value of Moran's *I* was 0.28; for the proteins with a significant *p*-value that did not meet log fold change threshold of ≥ 1 or ≤ -1, the median value of Moran's *I* was 0.10. For the proteins called as significant by the feature-independent analysis using the stringent $p = 0.001$ threshold, the median value of Moran's *I* was 0.33; in the case of proteins with a significant *q*-value ($q \leq 0.05$) that did not meet the $p = 0.001$ threshold, the median value of Moran's *I* was 0.23. These distributions show that the feature-independent analysis using Moran's captures part of the fold-change information of the feature-dependent analysis. However, the value of Moran's *I* does not distinguish between proteins with a large fold-change and a small fold-change. It only describes the spatial correlation, just as a Pearson correlation coefficient does not describe the magnitude of its calculated values. An advantage of the feature-dependent *t*-test method is that the fold-change can be used to filter data along with the *q* values and that a direction of the effect can be assigned (e.g. PYGL is upregulated in the solid tumour). This cannot be done with the feature-independent approach, and one negative effect is that the functional enrichment shown in cannot be performed. Indeed, performing the functional enrichment on the proteins significant by the feature-independent analysis does not return any significant pathways or gene ontology terms after FDR correction. One way to overcome this issue would be to cluster the data and then test for enrichment between clusters; this approach is further discussed in the next chapter.

3.3.2.2.1 Missing values and dispersion-like patterns impact detection by the feature-independent method

Under ideal conditions, it would be desirable for the feature-independent approach to capture all the proteins called as significant by the feature-dependent approach. However, this is not the case as 46 proteins are only called as significant by the feature-dependent approach. Further investigation of these proteins reveals two main reasons why these are 'missed' by the feature-independent analysis. The first is when the protein is not detected in the majority of either the solid tumour or the brain/tumour interface. If this protein is detected in at least two pixels within one of the tissue features, then *t*-test statistics can be calculated for that protein, and, in the case where the mean abundances differ between the features, the difference could reach significance. Moran's *I* values can also be calculated. However, the process of iteratively comparing a pixel's neighbourhood to the mean across the entire region will not detect spatial differential expression as the low number of low intensity pixels will not have a large effect on the neighbourhood. In addition, the information that a pixel is missing quantitative information (which could intuitively be used to infer difference) is not used. The second case is where a small area of a feature resembles that of dispersed data (high values surrounded by low values or vice versa), these will cause a drop in the value of Moran's *I* and may not have as large an effect on the *t*-test statistics.

One example of the missing value example is inactive dipeptidyl peptidase 10 (DPP10) (Figure 31A). DPP10 is detected in many solid tumour pixels but is generally missing within the brain/tumour interface region, with a few pixels along the margin showing low expression. DPP10 is called as significantly differentially expressed protein using the feature-dependent approach, with a $\log_2$ fold-change of +1.7 (solid tumour – brain/tumour interface), but a Moran's *I* of 0.033.

One example of a resemblance to dispersed data is EH domain-containing protein 4 (EHD4) (Figure 31B). EHD4 is quantified across the region of tissue sampled and has a $\log_2$ fold-change of +2.0 between the solid tumour and brain/tumour interface but a Moran's *I* of only 0.030. Across the tissue, there are a few pixels where very low EHD4 intensity values are surrounded by higher values, replacing two of these values (marked with a *) with the average value of their neighbours increases the value of Moran's *I* to 0.12 (p = 0.039) and replacing three other values (marked with a #) increases the value of Moran's *I* to 0.21 (p = 0.001).

This behaviour demonstrates a limitation of using a global indicator of spatial autocorrelation that describes the whole region with a single value. Occasionally, regions of dispersion-like spatial variation will hinder the detection of like-with-like regions clustering. While the example of EHD4 shows that removing these odd very low values increase the value of Moran's *I*, it should not be performed unless it can be determined to have been caused by a technical reason (e.g. failed digest).

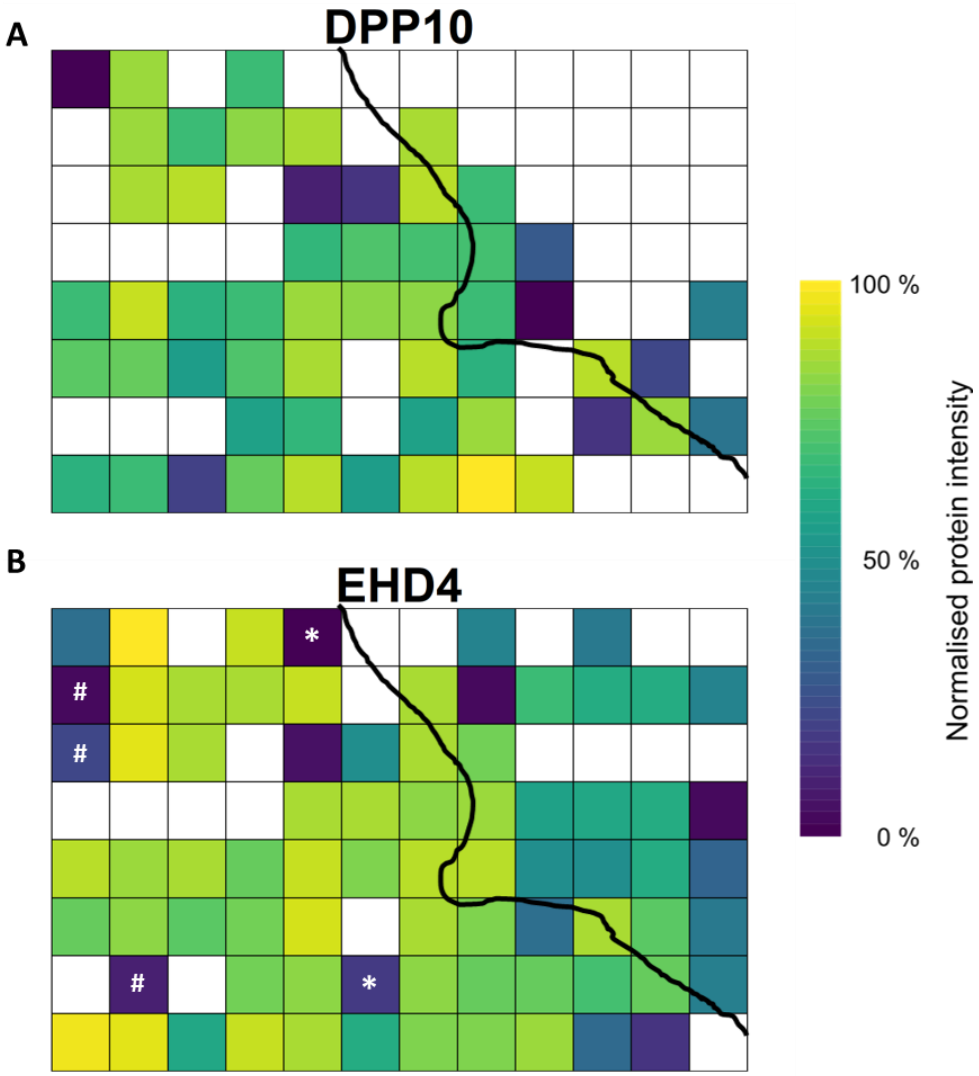

Figure 31 - Example of feature-dependent significant proteins 'missed' by the feature-independent method.

Normalised protein intensity of two example proteins mapped to the spatial positions of tissue isolated using LCM, each pixel represents one sample. (A) Dipeptidyl Peptidase Like 10 (DPP10) protein map showing missing expression in brain/tumour interface region. (B) EH Domain Containing 4 (EHD4) protein map showing dispersion-like features. Black line denotes the margin between the solid tumour (left of the line) and the brain/tumour interface (right of the line) as determined by the expert Neuropathologist. Normalised protein intensities are scaled separately for each protein. White = not quantified.

3.3.2.3 High-throughput spatial proteomics

The results described above were all taken from the Dionex/Lumos system analysis, which required ~8 days to acquire the 96 samples on 60-minute linear gradients with a cycle time of ~105 minutes. The same samples were also analysed on a high throughput system consisting of a state-of-the-art high-throughput LC, the EvoSep One, and a Bruker TimsTOF Pro with an 11.5-minute linear gradient and a 14.4 minute cycle time. This results in a throughput of 100 samples per day.

The EvoSep/TimsTOF system identified 12,750 peptides corresponding to 2,089 proteins in total, with a wide range of identifications per sample: 14 – 5,054 peptides corresponding to 8 – 1,327 proteins. Consistent with the Dionex/Lumos data, eight samples showed PEG contamination and were removed from the analysis. Here, none of the proteins identified were detected in every sample. It is expected and consistent with the literature that the EvoSep/TimsTOF approach identifies fewer proteins and is part of the inherent trade-off with the high-throughput analysis of a complex proteome, where the longer 60-minute gradients do not achieve even near-complete human proteome coverage. Of the proteins identified in the high-throughput data, only 3 % are not also found within the data generated by the Dionex/Lumos system (Figure S11). The same spatial pattern of proteins quantified per sample is visible as in Figure 26, indicating that this behaviour was a feature of the samples and not an issue with the LC-MS/MS data acquisition.

As with the Dionex/Lumos data, PCA shows a separation between the three main annotated tissue features of large blood vessels, solid tumour and brain/tumour interface (Figure 32A). This separation is not as strong, likely due to the PCA here being generated from fewer proteins than the Dionex/Lumos data. The data input to a principal component analysis cannot have missing values, so protein expression matrices were filtered to remove proteins that were not detected in at least 70 % of samples, and the remaining missing values imputed before PCA calculations. Therefore, the PCA for the Dionex/Lumos and EvoSep/TimsTOF data

were calculated from 1,918 and 385 proteins, respectively. Using this threshold resulted in discarding the majority of the data for the EvoSep/TimsTOF samples, a lower threshold of 50 % was also used for PCA calculations (Figure 32B). Using a threshold of 50 % resulted in 680 proteins' quantitative values being used for the PCA calculation. Using more of the data improves the separation of the solid tumour and brain/tumour interface samples.

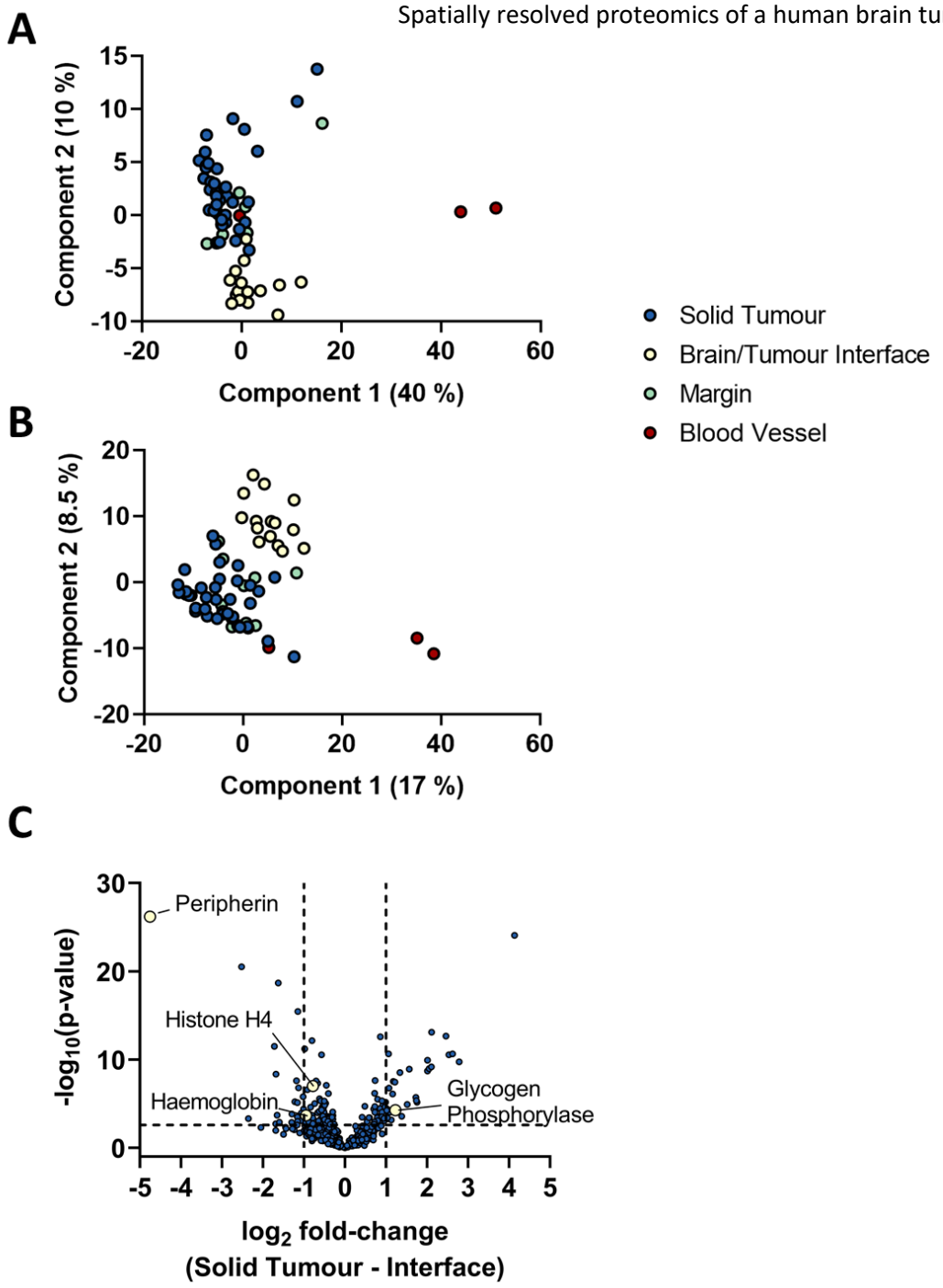

Figure 32 – Feature-dependent analysis of high-throughput ATRT spatial proteomic data

Analysis of ATRT spatial proteomic data from the EvoSep/TimsTOF high-throughput LC-MS/MS system. (A) Principal component analysis of ATRT tissue pixels using proteins quantified in at least 70 % of samples, with classification based on annotated histology image. (B) Principal component analysis of ATRT tissue pixels using proteins quantified in at least 50 % of samples, with classification based on annotated histology image. (C) Volcano plot of ATRT tissue pixels classified as wholly solid tumour or brain/tumour interface. Positive $\log_2$ fold-changes indicate higher protein abundance in solid tumour. The horizontal dashed line represents a 5 % FDR threshold. The vertical dashed lines represent ± 2-fold change.

Performing a two-sample *t*-test (plotted as a volcano plot in Figure 32 C) on the high throughput data shows good agreement between the two platforms, with 43 proteins called as significantly differentially expressed in both methods out of a total of 55 proteins called as significant with a $\log_2$ fold-change > ±1. From the *t*-test, 170 proteins have a q-value < 0.05 in the EvoSep/TimsTOF data and 165 of these were identified by the Dionex/Lumos system. 99 (60 %) of these proteins also have a *q*-value < 0.05 in the Dionex/Lumos data.

For an initial comparison between the two systems, I will again return to the four example proteins of peripherin, PYGL, haemoglobin and histone H4. Both peripherin ($\log_2$ fold-change = -4.8, $q < 0.001$) and PYGL ($\log_2$ fold-change = 1.3, $q = 0.0018$) show differential expression between the area of solid tumour and brain/tumour interface, as was the case in the Dionex/Lumos data, but both proteins have a less extreme $\log_2$ fold-change. In contrast to the Dionex/Lumos data, haemoglobin ($\log_2$ fold-change = 0.9 , $q = 0.007$) and histone H4 ($\log_2$ fold-change = 0.7, $q < 0.001$) also show differential expression between the areas of solid tumour and brain/tumour interface.

As the protein quantitation values are inferred based on measured peptide abundances, the peptides unique to histone H4 and haemoglobin were examined to investigate this difference. One peptide for histone H4 (VFLENVIR) displayed different behaviour between the two systems. In the Dionex/Lumos data, its normalised intensity was consistent across the areas of solid tumour and brain/tumour interface (*t*-test, *p*-value = 0.46). In the EvoSep/TimsTOF data, this peptide was significantly decreased in the solid tumour compared to the brain/tumour interface, with a $\log_2$ fold-change of -0.8 (*t*-test, *p*-value = 0.0081). This peptide's fold-change is similar to the protein's fold-change.

Similarly, one peptide for haemoglobin (GTFATLSELHCDKLHVDPENFR) was significantly decreased in the solid tumour compared to the brain/tumour interface, with a $\log_2$ fold-

change of -1.8 (*t*-test, *p*-value = 0.012). This peptide was not significantly different in the Dionex/Lumos data (*t*-test, *p*-value = 0.074).

For peripherin and PYGL, these proteins were only detected in pixels of brain/tumour interface or solid tumour, respectively. Therefore their fold-changes reported by the two-sample *t*-test are relative to the mean of the imputed values. These imputed values are sampled from the low-end of the intensity distribution, the mean of which will be above the detection limit. As these values were (mostly) non-missing in the Dionex/Lumos data, it is likely that here they are falling below the detection limit, and their replacement with imputed values will compress the $\log_2$ fold-change values. A schematic of this behaviour is presented in Figure 33. If the true difference is slightly larger than the chosen fold-change threshold, these compressed differences could result in a protein falling on either side of this threshold between different experiments. When comparing the $\log_2$ fold-changes of all proteins called as significant between the two platforms, the Pearson correlation coefficient is 0.90, indicating generally good agreement between the two platforms.

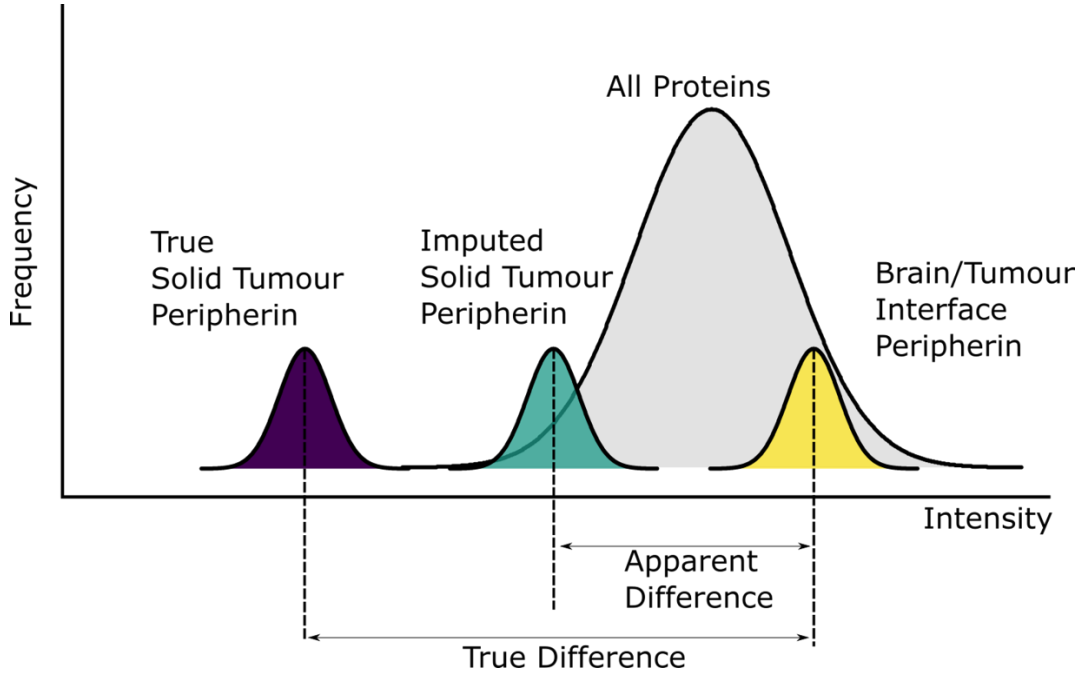

Figure 33 – Schematic of ratio compression caused by imputing missing values

A simulated example for peripherin of how imputing data that falls below the detection limit can distort the fold-changes between different sample types.

After analysing the high-throughput data using the feature-independent approach, 144 proteins have a *p*-value = 0.001, 289 have a *p*-value ≤ 0.01 and 565 have a *p*-value ≤ 0.05. After correcting for multiple testing, 245 proteins have an adjusted *p*-value ≤ 0.05. The four example proteins of peripherin, PYGL, haemoglobin and histone H4 have Moran's *I* values of 0.046, 0.33, 0.12 & 0.031, with adjusted *p*-values of 0.35, 0.049, 0.11 & 0.37. In contrast to the Dionex/Lumos data, these statistics do not correspond well to the feature-dependent approach.

Figure 34 shows the protein expression maps for peripherin, PYGL, haemoglobin and histone H4 derived from the EvoSep/TimsTOF data. These show general agreement with the maps from the Dionex/Lumos data. However, there are a much larger number of pixels with missing values in the high-throughput EvoSep/TimsTOF data. Across all proteins quantified, the data completeness is 19 %. Peripherin is quantified well in the brain/tumour interface but only in one pixel of the solid tumour. PYGL is only quantified in pixels of solid tumour and margin. As

discussed above, this exclusivity can hinder significance testing. In the feature-independent approach of testing for spatial autocorrelation, the exclusivity also presents difficulties as missing values are excluded from the global mean calculation, i.e. the mean that all spatial neighbourhoods are compared to for the detection of localised deviations from. If the pixels with quantified values are detected only in one spatial feature, then the global mean will be that of the feature, and no significant spatial autocorrelation will be detected. However, if the feature is large enough, it is likely to be clear from examining the map that there is some differential spatial expression of that protein, e.g. peripherin and PYGL in Figure 34. This explains the lack of correspondence to the feature-dependent approach for these proteins. This further highlights the impacts of missing values on the feature-independent analysis, suggesting that a minimum level of data completeness is required for this approach and that the use of this LC-MS/MS setup could be too optimistic for this approach.

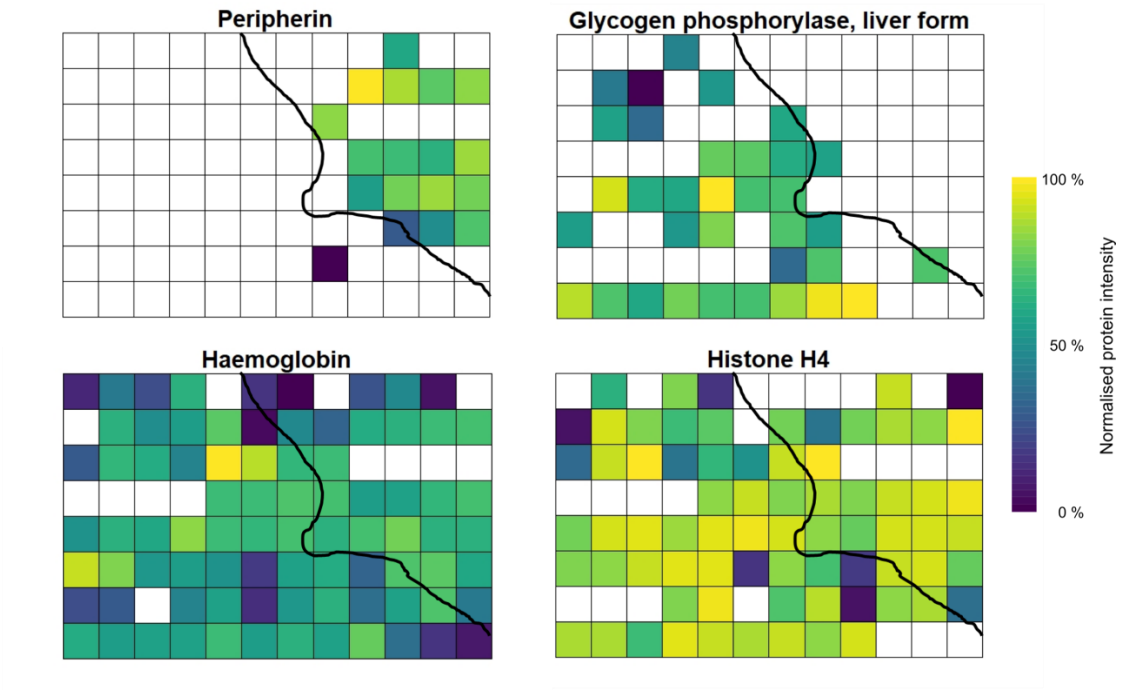

Figure 34 – High-throughput spatial proteomic maps of four example proteins

Normalised protein intensity of four example proteins mapped to the spatial positions of tissue isolated using LCM, each pixel represents one sample. The black line denotes the margin between the solid tumour (left of the line) and the brain/tumour interface (right of the line) as determined by the expert Neuropathologist. Normalised protein intensities are scaled separately for each protein. White = not quantified.

Table 3 – Summary of four example proteins between LC-MS/MS systems and statistical methods

	Low-throughput		High-throughput	
Protein	Feature-dependent	Feature-independent	Feature-dependent	Feature-independent
Peripherin	log_2 fold-change = -7.0 q < 0.001	Moran's I = 0.58 q = 0.011	log_2 fold-change = -4.8 q < 0.001	Moran's I = 0.046 q = 0.35
Glycogen Phosphorylase, Liver Form	log_2 fold-change = 4.5 q < 0.001	Moran's I = 0.41 q = 0.011	log_2 fold-change = 1.3 q = 0.0018	Moran's I = 0.33 q = 0.049
Haemoglobin	log_2 fold-change = -0.70 q = 0.14	Moran's I = 0.19 q = 0.019	log2 fold-change = 0.9 q = 0.007	Moran's I = 0.12 q = 0.11
Histone H4	log_2 fold-change = 0.18 q = 0.20	Moran's I = 0.09 q = 0.16	log_2 fold-change = 0.7 q < 0.001	Moran's I = 0.031 q = 0.37

3.3.3 Improving the success rate of tissue capture

When collecting large tissue areas, such as in the previous raster experiments or at the higher end of the response curve experiment (Figure 24), I noticed that the laser pressure catapulting (LPC) would occasionally fail to transfer the tissue into the cap. Often, the tissue would detach from the surrounding tissue but not move into the cap, requiring another LPC event or even multiple events to capture the tissue. This is undesirable as the LPC event ablates tissue. Usually, a small piece is left with this purpose in mind, but the repeated LPC events result in the ablation of part of the tissue area to be analysed – potentially losing information. In addition, monitoring and correcting these failures requires manual intervention, increasing the time taken to collect material. This behaviour was not noticed during the collection of material for the experiments where single cells and small tissue regions were used. In addition to the failed LPC events, during the sample transfer (in lysis buffer) from the cap to the tube, some tissue regions would remain behind within the cap requiring manual transfer using a pipette tip. Regions of the size isolated in the raster sampling are visible by eye due to the strong H&E staining, but I could not determine whether this manual transfer step caused the removal of some tissue – perhaps through adherence to the pipette tip during transfer. This behaviour could explain the multiple populations observed in Figure 26A, where the number of quantified proteins per sample displayed three clusters. Extra removal of tissue through repeated LPC events and the manual transfer would explain why some samples showed a lower number of quantified proteins, as lowering the input material would result in peptides dropping below the limit of quantification.

To try and remove this issue, an experiment was performed where 10^5 μm^2 (100,000 μm^2) square pixels of ATRT tissue were collected as 'subpixels' of different sizes and analysed using the optimised LCM-Proteomics method to determine if there is an optimal subpixel size to use in future raster experiments. Subpixel sizes started at 10^3 μm^2 (1,000 μm^2) and increased semi-logarithmically to 10^5 μm^2 (100,000 μm^2), representing a sample of all material collected in

one region. No attempts to correct failed LPC events were made. The mean (n = 3) number of proteins identified did not differ between the different subpixel sizes (One-way ANOVA, p = 0.608), but the variability between the triplicates was higher towards the more extreme ends of the subpixel range (Figure 35). One replicate of the 10^5 µm^2 (100,000 µm^2) samples is an outlier, with a number of proteins that would places it within the 'Low' cluster of the samples in the pilot ATRT experiment (Figure 26A), as this behaviour was not unique to one experiment or slide this indicates that there is a point of failure in the workflow for samples at least as large as 100,000 µm^2 when not collected in smaller areas. Based on these results, future raster experiments were collected using subpixels of approximately 10,000 µm^2 in area.

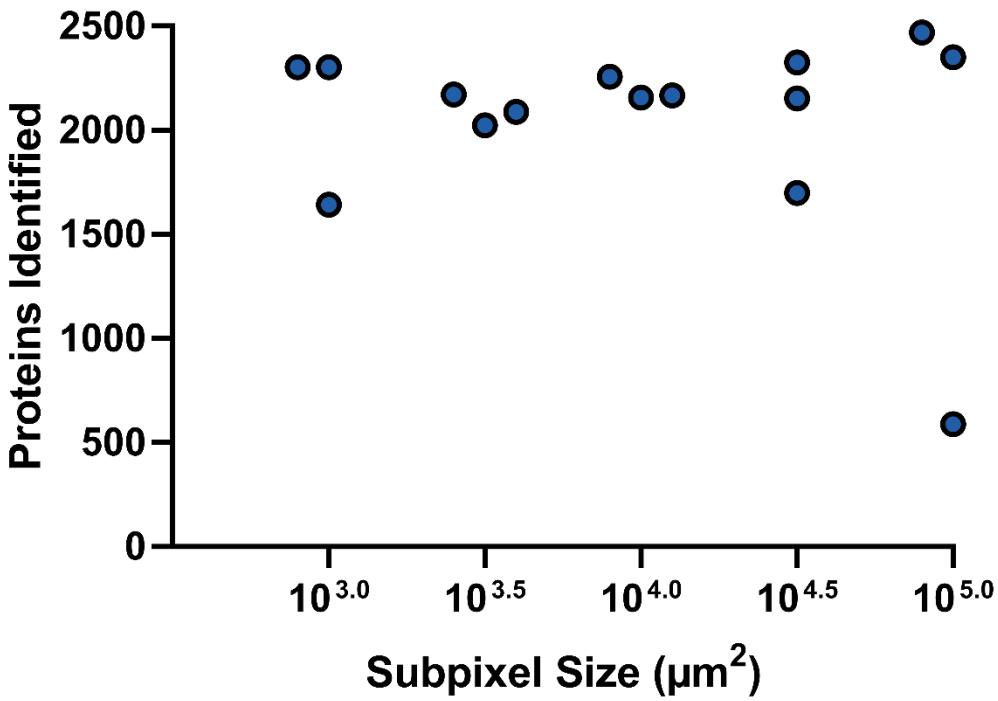

Figure 35 – Protein identifications from varying subpixel sizes

Protein identifications of 100,000 µm^2 of ATRT tissue collected in various sizes of smaller 'subpixel' regions. Samples for 10^5 µm^2 (100,000 µm^2) were collected as one whole square of tissue. Data points are jittered on the x-axis to avoid over-plotting of replicates.

3.3.4 Spatial proteomic analysis of a replicate ATRT section

After gathering the subpixel optimisation data, the raster sampling experiment was repeated on a slice from the same ATRT block as before (Figure 36A). Samples were collected in subpixels of 10,000 μm^2 from a 10 µm thick slice. Across the entire experiment, 3,994 proteins were quantified in at least one sample.

3.3.4.1 Feature-independent analysis

The 3,050 proteins that were quantified in at least nine samples were taken forward for feature-independent analysis. Of the proteins tested, 820 proteins have a *p*-value = 0.001, 1,173 proteins have a p-value ≤ 0.01 and 1,623 proteins have a p-value ≤ 0.05. After applying a Benjamini-Hochberg multiple testing correction, 1,375 proteins have an adjusted p-value ≤ 0.05. Returning to the four previous examples of peripherin, PYGL, haemoglobin and histone H4, these proteins have Moran's *I* values of 0.47, 0.61, 0.26, 0.22 and all have a *p*-value of 0.001 (0.0037 after false discovery rate correction).

The spatial expression pattern of the four example proteins is consistent with that of the initial raster experiment (Figure 36B). Peripherin has a higher expression level in the brain/tumour interface area, PYGL has a higher expression level in the area of solid tumour, and haemoglobin shows high intensity in the pixel that corresponds to the large prominent blood vessel. Histone H4 shows an even expression across the two annotated areas, but there is an area of seemingly lower expression in the region above the annotation line.

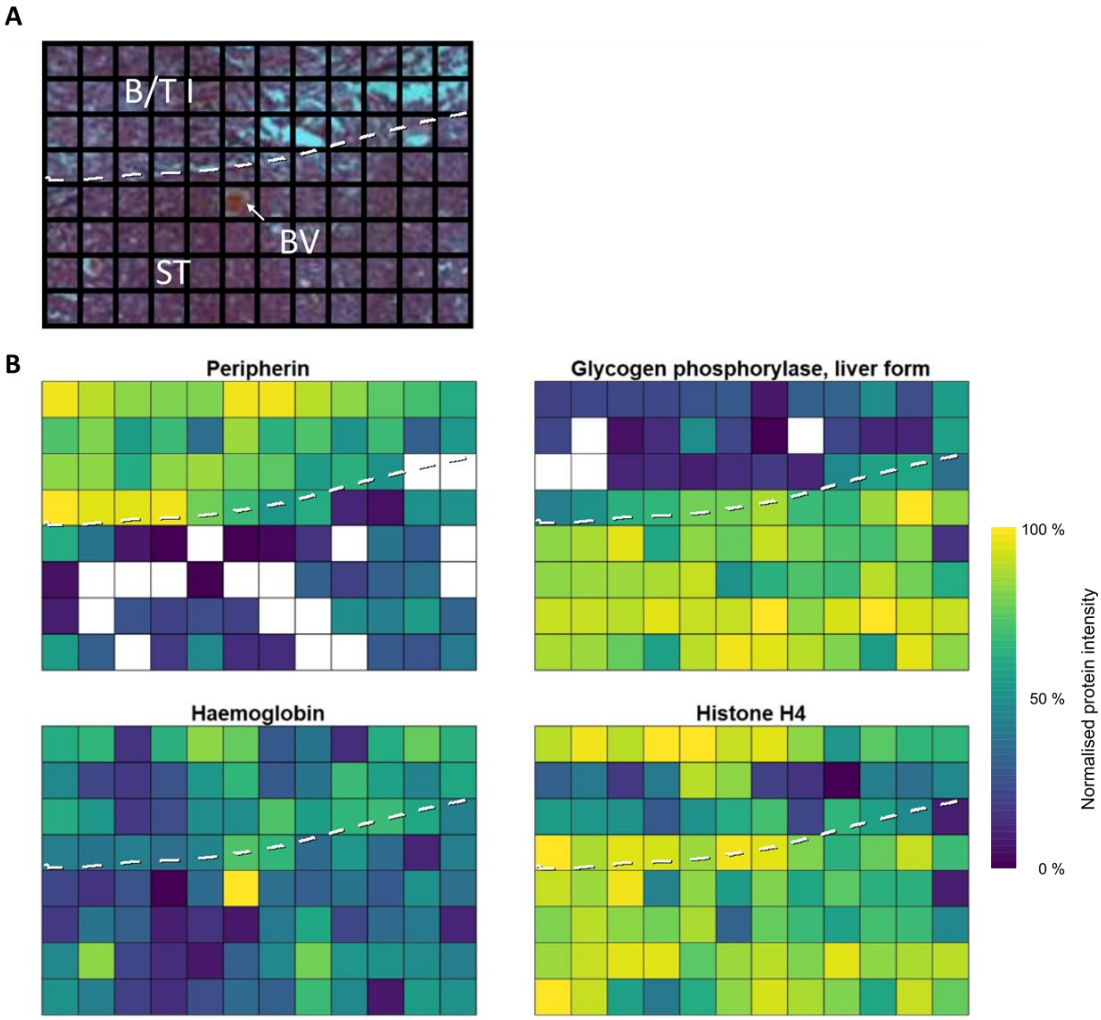

Figure 36 – Spatial proteomic maps of example proteins from a replicate tissue section

(A) Composite image of H&E stained ATRT section stitched from multiple fields of view acquired directly on the LCM instrument, with raster sampling LCM annotations overlaid. The dashed line represents the transfer of the pathologist's annotation to this section. BTI: brain/tumour interface. BV: blood vessel. ST: solid tumour. (B) Normalised protein intensity of four example proteins. Normalised protein intensities are scaled separately for each protein. The dashed line represents the transfer of the pathologist's annotation to this section.

To investigate why the expression of histone H4 displayed evidence for spatial expression patterns in contrast to the previous raster experiment, a Local Moran's *I* analysis was performed. The local analysis classifies individual pixels as 'high-high' (a high value in a neighbourhood of high values), 'low-low' (a low value in a neighbourhood of low values), two forms of spatial outliers: 'high-low' and 'low-high' (a high value in a neighbourhood of low

values and low values in a neighbourhood of high values, respectively) or pixels not significantly different to their neighbours. Here, the terms ‘high’ and ‘low’ are used with relation to the mean value across the entire protein map analysed and are not a comment on the absolute intensity.

The local analysis identified a cluster of pixels with low normalised protein intensity in the upper-right area of the expression of histone H4 (Figure 37A) that constitute an area enriched in ‘low-low’ pixels (Figure 37B). Referring back to an image collected at the time of sampling (Figure 36A & Figure 37C), this enrichment of low-low pixels corresponds to the centre of an area of the section with low tissue density.

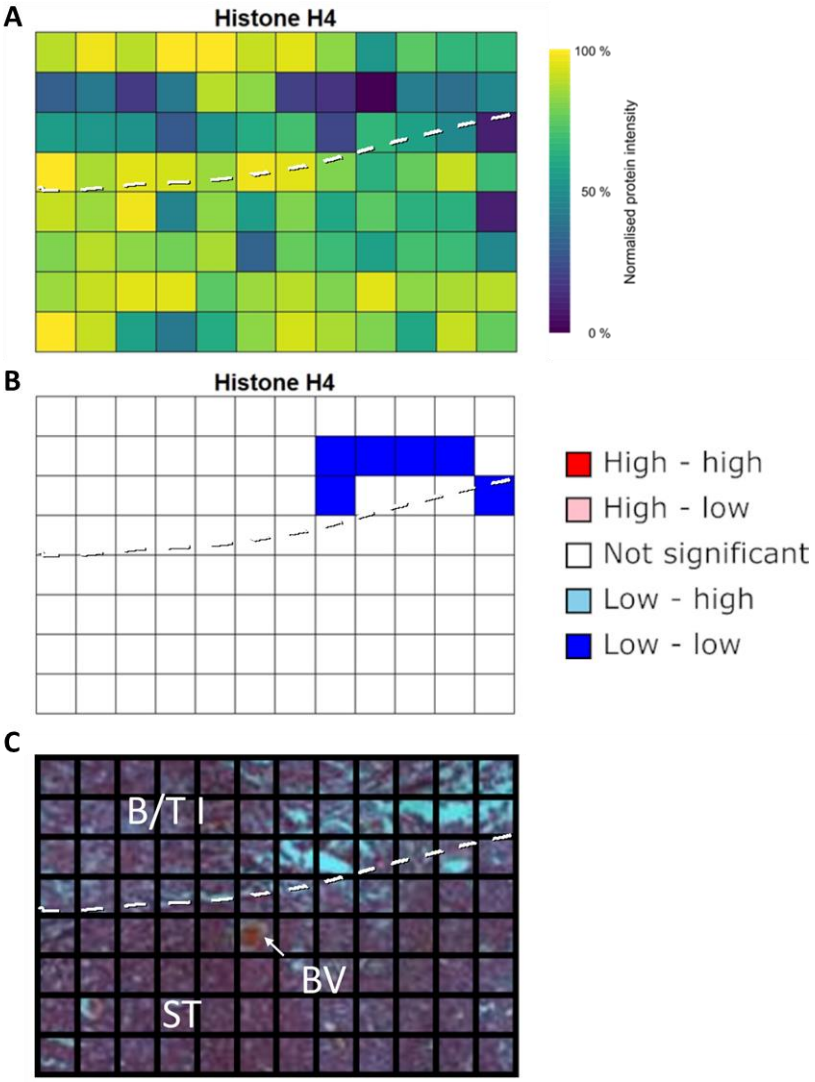

Figure 37 – Local Moran's *I* spatial autocorrelation analysis of Histone H4.

(A) Normalised protein intensity of Histone H4. (B) Pixel classification by local Moran's *I* analysis. (C) Composite image of H&E stained ATRT section stitched from multiple fields of view acquired directly on the LCM instrument, with raster sampling LCM annotations overlaid. The dashed line represents the transfer of the pathologist's annotation to this section. BTI: brain/tumour interface. BV: blood vessel. ST: solid tumour.

3.3.4.2 Feature-dependent analysis

A two-sample *t*-test identifies 134 proteins as differentially expressed between the areas annotated as brain/tumour interface and as solid tumour. In the area of solid tumour, 58 proteins were significantly increased and 76 proteins were significantly decreased compared to the area of brain/tumour interface using an adjusted *p*-value threshold of $q < 0.05$ and a

log_2 fold-change threshold of 1. Pixels where the annotation line passes through them were excluded from the analysis. Figure 38 shows a volcano plot of the *t*-test results with the four example proteins of peripherin, PYGL, haemoglobin and histone H4 highlighted. As before, both peripherin (log_2 fold-change = -6.3, $q < 0.001$) and PYGL (log_2 fold-change = 3.8, $q < 0.001$) show significant differential expression and histone H4 does not (log_2 fold-change = -0.3, q = 0.45). In contrast to the previous tissue section, haemoglobin also shows evidence for differential expression (log_2 fold-change = -1.6, $q = 0.029$).

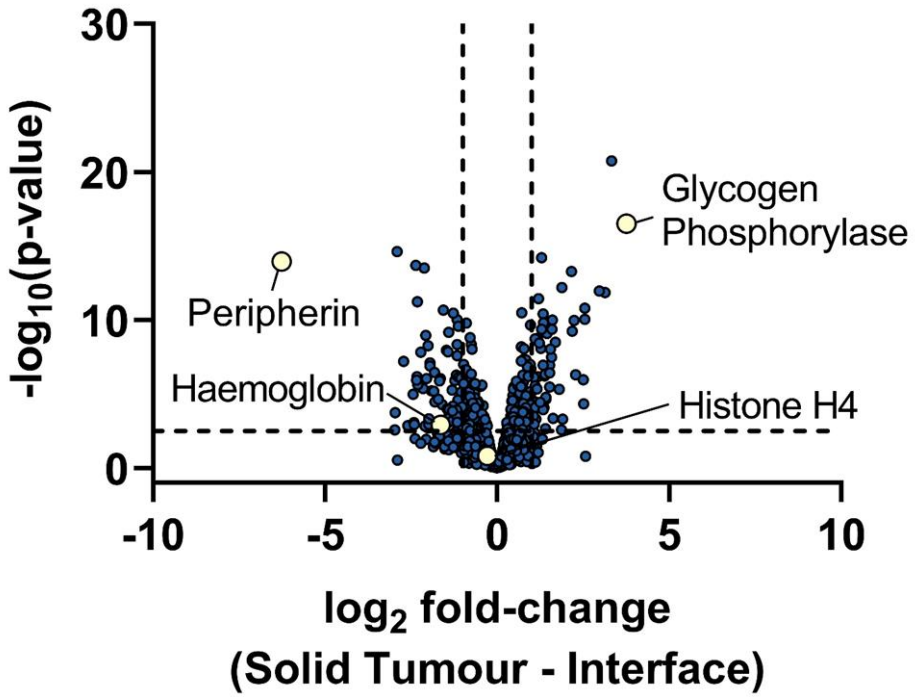

Figure 38 – Feature-dependent analysis of a replicate tissue section

Volcano plot of ATRT tissue pixels classified wholly as solid tumour or as brain/tumour interface. Positive log_2 fold-changes indicate the mean LFQ values is higher in solid tumour. The horizontal dashed line represents a 5 % FDR threshold. The vertical dashed lines represent ± 2 fold-change.

3.3.5 Comparison between pilot and replicate experiments

When comparing the two sets of significantly different proteins generated from the feature-dependent approach, 206 are called as significantly differentially expressed across both experiments. Of those 206 proteins, 117 (57 %) are called as significant in both samples. Figure

S12 shows the overlap both in identified protein groups, and significantly different proteins with a $\log_2$ fold-change of at least ± 1. 201 out of the 206 protein groups were identified and quantified in both experiments, demonstrating that lack of detection between experiments does not make a significant contribution to this low level of overlap.

3.3.5.1 Feature-dependent analysis

After inspection of proteomic maps of protein groups significant only in one experiment, four behaviour patterns are visible, examples of which are shown in Figure 39. The first is the spatial variation that is not captured by the pathologist's annotations. An example of this is the autophagy-related protein 3, which shows spatial variation across the annotated pathology in the first tissue section analysed but not the second.

In the second pattern, proteins are differentially expressed across the annotation in what essentially appears to be a presence/absence manner. The absence here could represent truly no expression of that protein, or it could be an apparent absence. An apparent absence would be due to the level of expression falling below the LC-MS system's detection and quantification limits. If a protein's expression is just below the detection limit in one of the annotated areas, then random variation within the tissue or analytical process could result in detecting that protein in a pixel. This needs to occur twice in the annotated area in one experiment and once, or not at all, in another experiment for the protein to be called significant in one experiment and be untestable in the other. An example of this pattern here is the neural cell adhesion molecule 2 protein (NCAM2). In both experiments, NCAM2 shows expression in the area of solid tumour. In the first experiment, NCAM2 is quantified in three pixels of the brain/tumour interface at a low level allowing for the calculation of *t* statistics and a $\log_2$ fold-change of 2.2. In the second experiment, it is only quantified in a single pixel.

The third pattern is where a protein shows more variability in one experiment in comparison to the other. This results in a non-significant *p*-value in the one experiment with more

variability. An example of this is transforming growth factor-beta-induced protein ig-h3 (TGFBI), which shows a higher level of variation within the second experiment's brain/tumour interface. As the two raster experiments were performed at different times, and other tissue sections were taken from the block between these experiments, the sections used for the two experiments were several hundred micrometres apart in the tissue block. The above example of ATG3 shows variation within the solid tumour on this length scale, and the variation seen in TGFBI, could be the same pattern but occurring along a different 3D axis.

The final pattern is where a protein shows spatial variation across the annotation in one experiment but consistent expression across the entire region in the other experiment. An example of this is the phosphoribosylaminoimidazole carboxylase and phosphoribosylaminoimidazole succinocarboxamide synthase enzyme, which shows significantly different expression in the first experiment but consistent expression across the entire region in the second experiment.

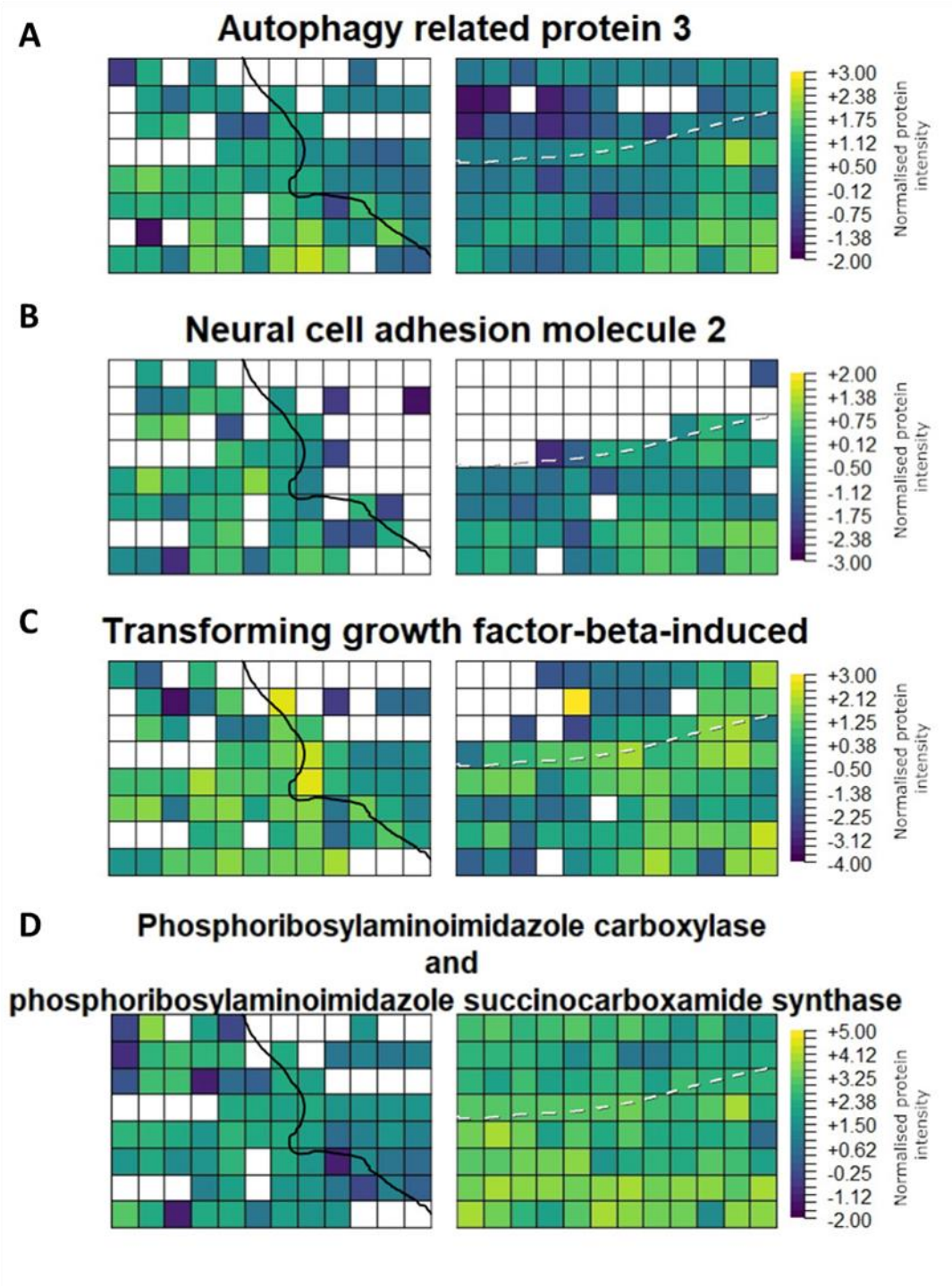

Figure 39 – Spatial proteomic maps of example proteins with different behaviour between experiments

(A) Spatial variation that is not captured by the pathologist's annotations. (B) Presence/absence expression pattern. (C) Different variability in brain/tumour interface between experiments. (D) Spatial variation in one experiment and not the other.
The black line denotes the margin between solid tumour (left of the line) and brain/tumour interface (right of the line) as determined by the expert Neuropathologist. The White dashed line represents the transfer of these annotations to this slide. White = not quantified.

3.3.5.2 Feature-independent analysis

After comparing the performance of the feature-dependent approach between replicate experiments, the feature-independent approach was also compared. Before comparison, the data was filtered for proteins detected in at least 8 pixels in each experiment. From the two experiments, 431 proteins and 1,275 proteins have Moran's I *q*-values < 0.05, with an overlap of 249 proteins (57 % of the proteins in the first experiment) (Figure 40). The second experiment has many more proteins called as significant (using both $q < 0.05$, and $p = 0.001$ as a metric), this could be caused by several reasons (outlier pixels, different tissue density, and missing values) and are discussed below.

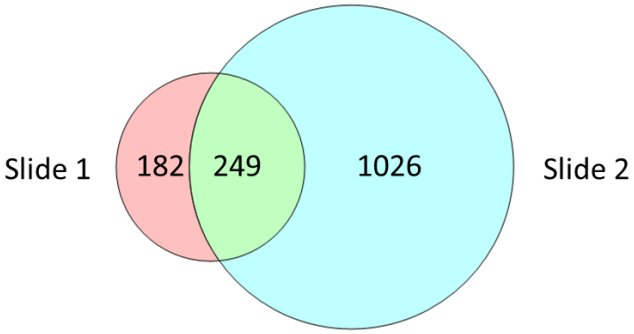

Figure 40 – Overlap of spatially significantly differentially expressed proteins using the feature-independent method

Overlap of proteins called as significant by the feature-independent approach using Moran's test of spatial autocorrelation with an FDR of 5 % between two replicate experiments.

Firstly, the original experiment has some apparent outlier pixels with a low number of proteins identified, which causes many missing values within the maps and could be suppressing the values of Moran's *I*, as discussed in the "Missing values and dispersion-like patterns impact detection by the feature-independent method" section above. The data completeness of slide one was 46 %, this increased to 53 % in the data from slide two. Comparing the distribution of quantified values per protein shows a shift towards more complete data in the replicate slide (Figure 41). Many more proteins having a complete or near-complete map is likely contributing to this increase in spatially significantly differentially expressed proteins.

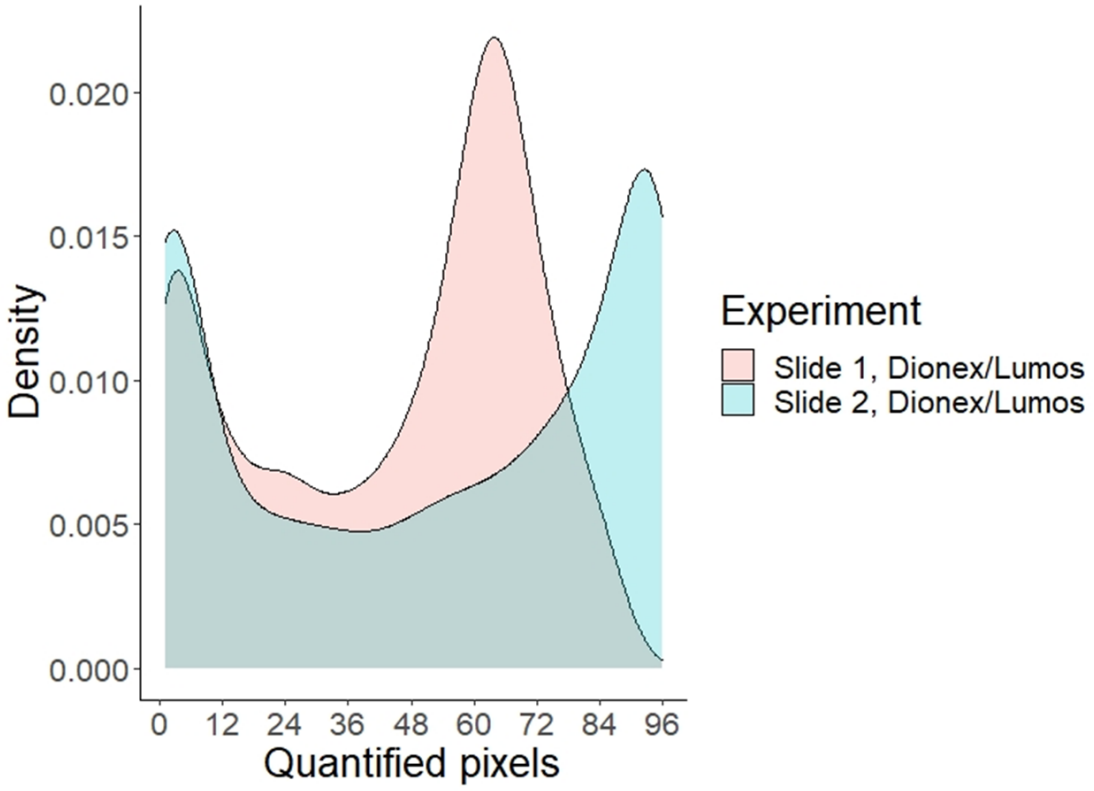

Figure 41 – Density plots of quantified pixels per protein map

Density plots of quantified pixels per protein map between the pilot and replicate ATRT tissue sections

Next, the second experiment has a region of apparently lower tissue density, indicated by the H&E image and histone H4 abundance. As shown in Figure 37, this area overlaps with low histone H4 abundance, driving a larger and significant value of Moran's *I* for this protein. Examining the local Moran's *I* pixel classifications of the 1,026 proteins uniquely significant in the second experiment reveals many proteins that also display a significant area of low protein expression (examples of which are shown in Figure S13).

When comparing the proteomic maps from both tissue sections for those proteins called as significant only in the first experiment, a feature common to some of these proteins is the presence of missing values in one or both experiments (Figure 42). In the case of activated leukocyte cell adhesion molecule, there are more pixels quantified in the first experiment that form a contiguous region displaying spatial variation. There are more missing values in the

corresponding map from the second experiment, and those pixels where it is still quantified are less connected.

For cadherin 11, this protein is only quantified in 10 and 12 pixels per map. In the maps for both slides, there are two unconnected neighbourhoods. In the first slide, one neighbourhood has a higher protein abundance than the other, and both are within the brain/tumour interface region. As each neighbourhood is contiguous, the Moran's *I* can be calculated, and it detects the deviation of these neighbourhoods from the global mean of that map. For the second slide, both of the neighbourhoods have similar protein abundances resulting in no deviation from the overall mean and a call of non-significant by the Moran's *I* test. Under the assumption that the true protein abundances in the pixels with missing values are below the detection limit of the LC-MS/MS system, then by eye, it would be reasonable to assume that the variation is also present in the second slide. However, this variation is being masked from the statistical test due to the missing values, as the Moran's *I* test does not use the information on missingness.

For the map of dystroglycan generated from the first slide, two contiguous neighbourhoods show spatial variation – one is within the brain/tumour interface region, and the other is contained within the solid tumour. There are many missing values in the map of the second slice, and the protein abundance is, in general, lower in this slice. Of those pixels where the protein was quantified, the only contiguous region is within the brain/tumour interface. A general decrease in intensity for this protein could be responsible for this, perhaps due to sampling a different region within the tissue block or experimental variation.

Stromal cell-derived factor 4 was quantified across the solid tumour in both slides but only the brain/tumour interface region within the first slide. Again, this allows the Moran's *I* test to detect the spatial variation within the data from the first slide, but not the second, as the missing values from the brain/tumour interface area are not used.

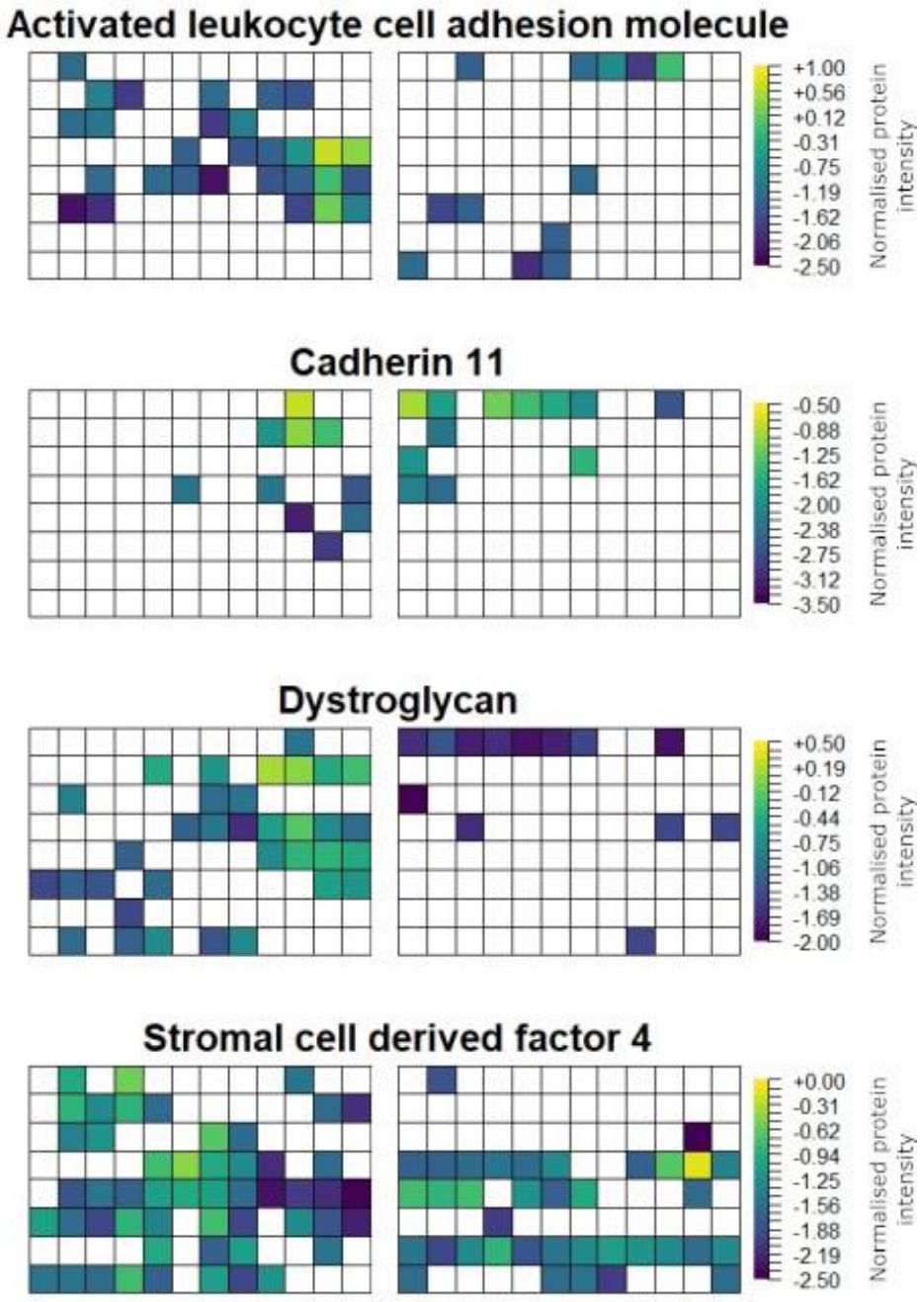

Figure 42 - Spatial proteomic maps of example proteins significant by feature-independent analysis in experiment one only

Normalised protein intensity maps of four example proteins called as significant by the feature-independent Moran's *I* analysis only in the first experiment (left maps) and not the second experiment (right maps). White = not quantified.

From these data, it is clear that the presence and distribution of missing values affect the detection of proteins displaying spatial variation within the sampled regions and when

comparing regions sampled from within the same tissue. This limitation could be addressed in the future by collecting deeper, more complete proteomes for each pixel and applying spatially aware data imputation methods. The recent improvements in liquid chromatography and mass spectrometry instrumentation and data acquisition methods for data-independent analysis (DIA) allowed Brunner *et al.* to quantify over 3,000 proteins from 1 ng of input material with only 6 % missing values between replicate analysis of HeLa lysate[250]. When applying their methods to single HeLa cells, the mean protein completeness per cell was 54 %. In the two spatial proteomics experiments presented within this chapter, the mean data completeness was 49 % (46 % and 53 % completeness per experiment). As DIA analysis generally results in fewer missing values than the DDA approach used in this chapter, I would expect an increase in the data completeness if these spatial proteomics experiments were acquired using DIA. The use of DIA could be particularly beneficial to the high-throughput approach that only has a data completeness of 19 %. In addition, DIA acquisition methods can quantify more proteins than DDA methods within the same amount of time. Even though DIA increases data completeness by removing missing values and generating deeper data, missing values will still occur when the protein is genuinely not present or lies below the, albeit lower, detection limit. Because of this, improvements in the downstream analysis of the spatial proteomic maps will also be required. The handling of missing values in proteomics is far from a solved problem, and there are a wide range of methods with different performance between different proteomic data modalities[251–255]. In the context of spatially resolved proteomics, methods that use the spatial relationship between samples would presumably perform well. For example, spatially aware imputation methods have shown promise for analysing spatial transcriptomics data by increasing the accuracy of imputation when accounting for the spatial relationships of samples [256].

3.4 Conclusion

Here, I have demonstrated that the optimised workflow developed in Chapter 2 can be applied to fresh-frozen tumour tissue in a sensitive, robust and spatially resolved manner. The method shows a clear signal from a starting material of ~25-50 cells isolated by LCM from ATRT tissue, equivalent to a resolution of 50 μm^2 and can map over 3,000 proteins at a resolution of 350 µm. The data generated can distinguish between broad histopathological features using unsupervised and supervised statistics, and integrating this data to the pathway level reveals differences generally consistent with tumour biology. Analysis of the data within the spatial context highlights proteins showing spatially variable abundance within the tissue without the need for prior knowledge of the tissue composition through spatially aware statistical methods. In some circumstances, missing values negatively impact the data and represent an avenue for further data acquisition and analysis improvements. The spatially resolved proteomics approach described here has the potential to generate proteomic maps in many contexts to aid research into protein expression within heterogeneous tissues.

4 Further improvement, application and validation of spatially resolved proteomics

4.1 Introduction

In the previous chapter, the optimised method developed in Chapter 2 was applied to perform spatially resolved proteomic measurements of ATRT tumour tissue. Although proteomic depth is compromised in high-throughput experiments, the method performs well, reflecting known aspects of tumour biology. In addition, the investigation of pathway-level changes required prior knowledge of tissue structure. The analysis of samples in a low-thoughput manner will limit both the number of replicate samples and their number of pixels in a given time. Therefore, higher-throughput methods which do not compromise on proteomic depth are needed. In this chapter, further improvements to the sample preparation, LC-MS system configuration and data analysis methods are described, along with orthogonal validation efforts. Here it is shown that short gradient nLC-MS/MS on a fast scanning instrument offers improvements in throughput without compromise in data depth, allowing for the analysis of larger regions within the same analysis time.

4.2 Materials and Methods

Reagents were purchased from Sigma Aldrich unless otherwise specified. Solvents for LC-MS use were LC-MS grade and purchased from Millipore. Formic acid was purchased from Fisher Scientific.

Tissue preparation, LCM and SP3 proteomic sample preparation were performed as per Chapter 3 from the same tissue block with the following modifications. The LCM sample collection system was upgraded to allow the collection of samples into a 96-well format (RoboMover, Zeiss). Samples were collected into 20 µL RIPA buffer (Pierce, #89900) in either the cap of 200 µL PCR tubes or flat PCR cap-strips. Collected samples were immediately placed in dry ice (tube caps) or onto a 96-well PCR plate pre-cooled in dry ice (cap-strips).

4.2.1 LC-MS/MS

The Dionex/Lumos system was operated with the same configuration and parameters as Chapter 3. The EvoSep/TimsTOF system was also operated with the same configuration and parameters as Chapter 3, with the exception of parameters being optimised which are described in the text.

Dionex/TimsTOF:

Peptides were analysed by nLC-MS/MS using a Dionex Ultimate 3000 (Thermo) coupled to a timsTOF Pro (Bruker) using a 75 µm x 150 mm C18 column with 1.6 µm particles (IonOpticks) at a flow rate of 400 nL/min. A 17-minute linear gradient from 2 % buffer B to 30 % buffer B (A: 0.1 % formic acid in water. B: 0.1 % formic acid in acetonitrile). Sample injection was performed using direct loading onto the analytical column via a 5 µL sample loop. The TimsTOF Pro was operated in PASEF mode. The TIMS accumulation and ramp times were set to 100 ms, and mass spectra were recorded from 100 – 1700 m/z, with a 0.85 – 1.30 Vs/cm^2 ion mobility range. Precursors were selected for fragmentation from an area of the full TIMS-MS scan that excludes most ions with a charge state of 1+. Those selected precursors were isolated with an ion mobility dependent collision energy, increasing linearly from 27 – 45 eV over the ion mobility range. Three PASEF MS/MS scans were collected per full TIMS-MS scan, giving a duty cycle of 0.53 s. Ions were included in the PASEF MS/MS scan if they met the target intensity threshold of 2000 and were sampled multiple times until a summed target intensity of 10000 was reached. A dynamic exclusion window of 0.015 m/z by 0.015 Vs/cm2 was used, and sampled ions were excluded from reanalysis for 24 seconds.

4.2.2 Data analysis

4.2.2.1 Protein identification and label-free quantitation

Protein identification and quantification of raw mass spectrometry data files were performed using MaxQuant as described in Chapter 3. For the analysis of histone trimethylation, a

variable modification of +42.0470 Da was allowed on lysine residues, with a maximum of 3 missed cleavages.

4.2.2.2 Experimental design and analysis of mass spectrometry parameter optimisation

Response surface methodology experiments were performed using a central composite design with 13 experimental points to investigate linear and quadratic effects and interaction effects. Each parameter was varied across five levels and these levels are shown in (Figure S15 and Figure S16). The central point was repeated 5 times to measure variability. Collected data were analysed using response surface regression within Minitab 19 (Minitab, Inc.). Response surfaces are plotted using the fitted model.

4.2.2.3 Feature-independent analysis

Feature independent analysis of protein intensity distribution was performed as described in Chapter 3. Empty pixels or those covering the large region of haemorrhage at the tissue's edge were not included in these calculations.

4.2.2.4 Clustering

Empty pixels and pixels covering the large region of haemorrhage were not included. A distance matrix was built containing the Euclidean distance between each pixel's set of protein LFQ values. Hierarchical clustering of the distance matrix was performed in R using the "average" agglomeration method. Dendrograms were cut into clusters based on dendrogram height alone (R's stats::cutree function), or using the Dynamic Tree Cut method using the "hybrid" method with a minimum cluster size of 1[257].

A complete input matrix is required for UMAP visualisation, so proteins with fewer than 70 % valid values across the experiment were removed. The remaining missing values were imputed in Perseus on a per-sample basis by random draws from a normal distribution using a width of 0.3 and a downshift of 1.8. UMAP dimensionality reduction was performed on this imputed data with default settings, and the first two embedding components plotted and samples coloured according to their cluster assigned by Dynamic Tree Cut at a height of 100.

4.2.2.5 Pathway analysis

A one-way ANOVA test was performed in R to test for differences in means between the clusters generated by Dynamic Tree Cut at a height of 100. A pairwise post-hoc correction was applied using Tukey's Honestly Significant Difference method. The resulting pairwise comparisons were used as inputs to ClusterProfiler's gene set enrichment analysis to test for enrichment of Gene Ontology Biological Process terms using a Fisher's exact test with a 5 % FDR threshold [219].

4.2.2.6 Cell-type map generation

Protein identifiers were converted to gene symbols and single sample gene set analysis (ssGSEA) was performed with default parameters using the "ssgsea-gui.R" script at https://github.com/broadinstitute/ssGSEA2.0 to generate cell-type enrichment scores based on curated cell-type marker proteins from the Molecular Signatures Database (MSigDB) [258–260]. The enrichment scores for the cell type within each sample were plotted in their spatial context.

4.2.3 MALDI Imaging

4.2.3.1 Trypsin digestion

The vacuum dried sections were washed according to a Carnoy procedure as described in Deutskens *et al.* [261] and dried again under vacuum for at least 30 min. The slides were scanned with a TissueScout scanner at 3200 dpi (Bruker) to generate a reference image for later position teaching. Trypsin was applied using a TM-sprayer (HTX Technologies) as described in Ly *et al.*[262] and digested for 2 h at 50 °C.

4.2.3.2 MALDI matrix preparation

For lipid imaging, dihydroxybenzoic acid (DHB) was dissolved at a concentration of 15 mg/ml in 90 % ACN, 0.1 % TFA and sprayed on top of the ATRT sections using a TM-sprayer (HTX Technologies). The matrix was applied in a criss-cross pattern with 3 mm track spacing at a 1200 mm/min nozzle velocity. Fourteen layers were sprayed at a 0.125 ml/ml flow rate using

50 % ACN as the liquid phase at 10 psi pressure. The nozzle temperature was set to 60 °C and the distance of the nozzle to the section was 40 mm.

For peptide imaging, slides were covered with alpha-cyano-4-hydroxy-cinnamic acid (CHCA) matrix as described[262].

4.2.3.3 MALDI Imaging data acquisition

All imaging data were acquired on a timsTOF fleX instrument (Bruker) which is equipped with a dual ESI and MALDI source in positive Q-TOF mode. External calibration was performed using red phosphorous which was spotted next to the section.

For lipid imaging, the laser was operated in beam scan mode, ablating an area of 15x15 µm resulting in a pixel size of 20 µm. The repetition rate of the laser was set to 10 kHz and 400 laser shots were acquired per pixel. Data were acquired in the mass range 300-1400 m/z.

Peptide imaging data were acquired with at 50 µm pixel size in the mass range 600-3400 m/z. 1600 laser shots were accumulated per pixel at a repetition rate of 10 kHz. These MALDI imaging experiments were kindly performed by Dr Janina Oetjen.

4.2.3.4 MALDI Imaging data analysis

The software SCiLS Lab (version 2020a; Bruker) was used for MALDI Imaging data analysis. All data were root-mean-square (RMS) normalised. After importing the data, an unsupervised segmentation was calculated using the bisecting k-means algorithm and a peak list containing 302 m/z-intervals for the lipid imaging data and 498 m/z-intervals for the peptide imaging data, with correlation used as a distance metric.

The resulting segmentation map was split into several clusters that resemble the histopathology of the tumour section, and regions of molecular similarity were generated. Finally, an area correlating to the tumour was defined by comparing the cluster regions (segments) with the histopathological examination of a serial section stained with H&E.

The co-localisation tool of the SCiLS Lab software was employed to detect m/z-features that are overrepresented in the tumour area defined by the unsupervised segmentation approach. This tool relies on the Pearson correlation and can be used to find similarities between the m/z-images and the region mask represented by the defined tumour area. A correlation factor of 1 means that region mask and m/z-image match perfectly, while a factor of 0 indicates no correlation. Anticorrelated images would have a correlation factor of -1.

4.2.4 Immunohistochemistry

Sections were cut as above and were mounted to superfrost glass slides for IHC and air-dried. Slides were fixed in ice-cold acetone for 10 minutes, washed twice with TBS/T (20 mM Tris, 150 mM NaCl, 0.05 % Tween 20) and blocked with blocking solution (5 % goat serum in TBS/T) for 60 minutes at room temperature. Primary antibodies diluted in blocking solution and incubated at RT for 60 minutes or 4 °C overnight. Sections were washed three times with TBS/T. Staining visualisation was performed by incubating with a cocktail of anti-mouse and anti-rabbit secondary antibodies conjugated to horseradish peroxidase (Envision Kit, Agilent) for 60 minutes at room temperature. Sections were then washed with TBS/T three times and incubated with 2 % 3,3'-diaminobenzidine for 5 minutes, immersed in water and then counterstained with Harris' Haematoxylin for 1 minute. These immunohistochemistry experiments were kindly performed by Connor Scott.

Primary antibodies used and dilutions: rabbit anti-PYGL, 1:100, 4 °C overnight, HPA000962 (Atlas Antibodies); rabbit anti-ASPH, 1:1000, RT 60 minutes, NBP2-34125 (Novus Biologicals); mouse anti-CD45 (PD7/26 + 2B11), 1:200, RT 60 minutes, ab781 (Abcam).

4.2.5 Public data acquisition

The "Nanowell_Imaging_Uterus/StromaDominant/MatchBetweenRuns/proteinGroups.txt" file was downloaded from the Mass Spectrometry Interactive Virtual Environment (MassIVE; [https://massive.ucsd.edu]) repository with accession number MSV000084421 [139].

4.3 Results and Discussion

4.3.1 Adaption of the existing protocol to 96 well plates

The method optimisation work on the LCM-Proteomics workflow was designed for lower-throughput applications such as comparing a few cell types, so tissue was captured into and processed in individual PCR tubes. To more efficiently and consistently process the number of samples planned for the spatial proteomics experiments, the workflow required scaling to a format more compatible with working with 96-well plates in a high throughput fashion. During the time that the work for this chapter was being performed, the laser capture microscope was upgraded from a capacity of collecting two samples at one time (i.e., the capture device could hold two caps for collection) to capable of collecting 96 samples at one time (i.e., the capture device could hold 96 caps for collection). Several modifications were required: Collection of samples into strips of PCR caps rather than individual tube lids; transfer of samples to a PCR plate rather than the tube body of the cap it was collected into; use of a different magnet compatible with 96-well plates which pulls beads to a lower position in the well; Resuspension of beads by pipette rather than water-bath sonication. Due to this array of changes, a repeat titration was performed as in Figure 24 and analysed on the Dionex/Lumos system. Figure 43 shows a comparison of this titration and the one performed in Figure 24; the same logistic curve-like response can be seen, demonstrating that the method performs equivalently.

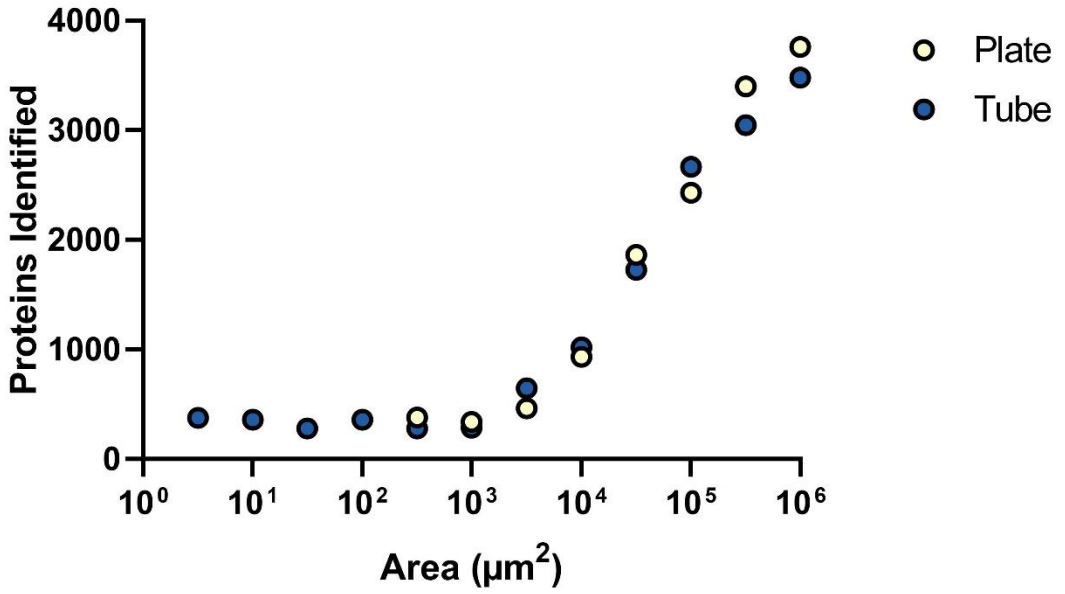

Figure 43 – Protein identified from a titration of ATRT area after high-throughput sample collection and preparation modifications

Number of proteins identified from titration of ATRT area on the Dionex/Lumos system with sample collection and processing in a 96 well plate (Plate) vs individual PCR tubes (Tube).

4.3.2 High throughput LC-MS/MS optimisation

To determine whether the default instrument settings for the TimsTOF Pro were optimal on our instrument at lower peptide loads, several Design of Experiments approaches were used to analyse 50 ng of HeLa peptide standard on the EvoSep One's "100 samples per day" method. First, a screen of the instrument parameters was performed to determine which parameters significantly affect the number of identified peptides. This screen identified the accumulation time and ramp time of ions in the TIMS device, and the intensity threshold and target intensity of ions chosen for PASEF MS/MS fragmentation as parameters that could benefit from optimisation.

4.3.2.1 Response surface optimisation of TIMS accumulation and ramp time

A response surface optimisation experiment was performed to investigate the parameter space of the ion accumulation and elution time. Nine combinations of accumulation and ramp times were tested using a central composite design to detect any curvature in the response.

The choice of these parameter settings was based on the screen and are shown in Figure S15. The model's output can be intuitively visualised as a contour plot, where the x and y axes represent the independent variables (here, accumulation and ramp time) and the 'height'/colour represents the predicted response (here, identified peptides). Figure 44A shows that the predicted optimal values lie within a region encompassing the default setting of 100 ms for both accumulation and ramp time, and also suggests that higher values could be beneficial. However, the interaction plot in Figure 44B shows that both the 100 ms and 120 ms levels have the same upper bound of ~6,400 peptides identified. In addition, the 100 ms / 100 ms data has a higher number of PASEF MS/MS scans per peptide in comparison to the 120 ms / 120 ms data (3.5 and 3.1 respectively). Based on this, an accumulation time of 100 ms and ramp time of 100 ms were kept for future samples.

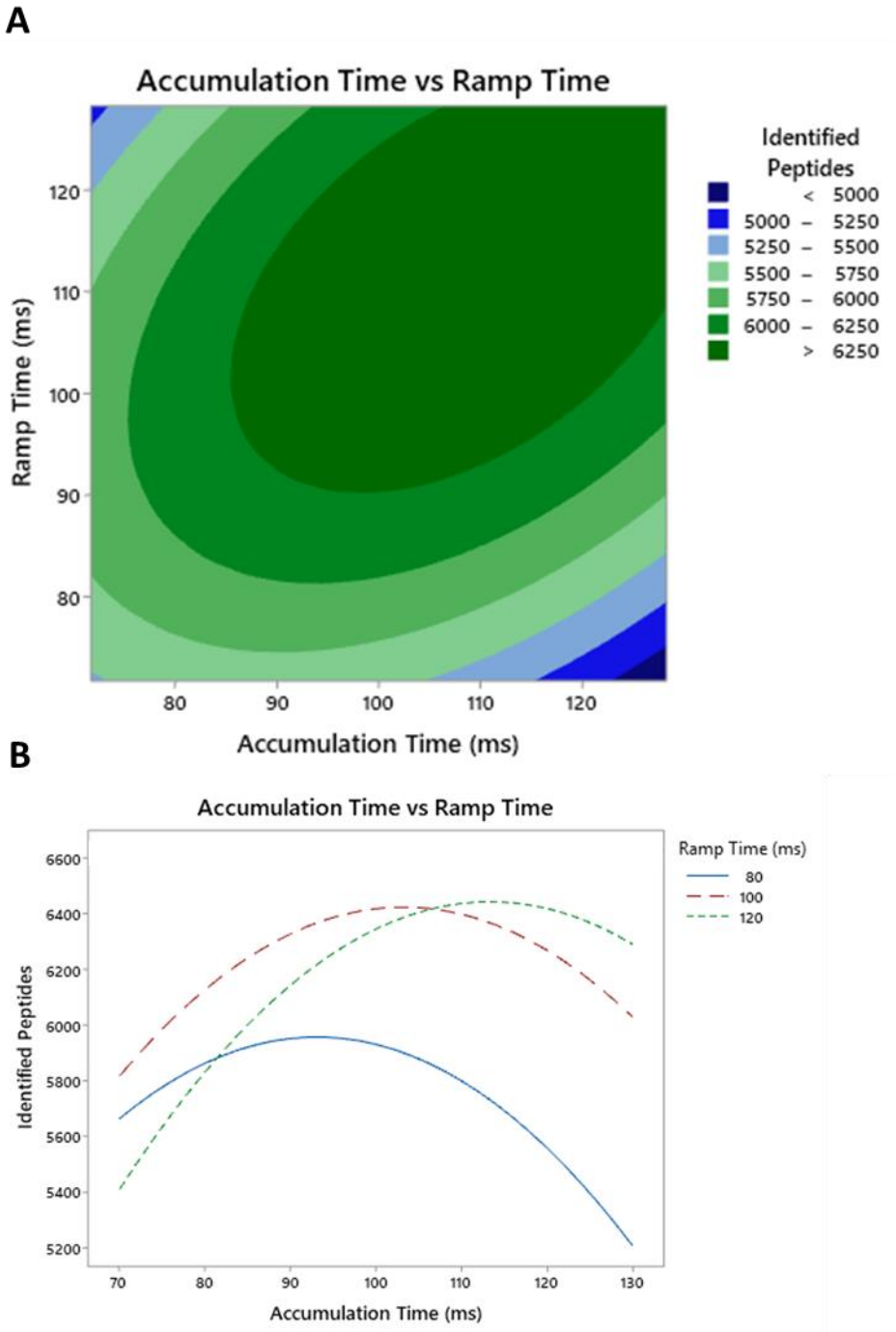

Figure 44 – Optimisation of instrument ion mobility parameters

(A) Contour plot visualisation of response surface for the number of identified peptides when varying the TIMS accumulation and ramp times. (B) Interaction between TIMS accumulation and ramp times showing that optimal value of accumulation time depends on the ramp time

4.3.2.2 Response surface optimisation of intensity threshold and target

The parameter space for the PASEF MS/MS ion intensity target and intensity threshold was investigated in the same fashion as above and the parameter combinations tested are shown

in Figure S16. The default parameters are 24,000 for the target intensity and 2,000 for the intensity threshold. Figure 45A shows that the target intensity value has a larger effect on the number of identified peptides than the intensity threshold, with a lower target increasing peptide identifications. Figure 45B shows that the intensity threshold has a larger effect at lower target intensities, with a threshold of 2,000 identifying the most peptides. The parameter combination with the highest number of identified peptides was a target intensity of 10,000 and an intensity threshold of 2,000. This combination resulted in a mean of 7,793 identified peptides (n=2). When operating in PASEF mode, the instrument will target the same precursor for fragmentation in multiple PASEF MS/MS scans until the target intensity is reached. These reconstructed spectra originating from the same precursor are combined to give a single spectrum for identification. It, therefore, follows that a reduction in target intensity would allow for more precursors to be targeted for fragmentation.

However, a benefit of sampling the same precursor multiple times and creating a summed spectrum is that it can increase the quality of the spectrum by increasing the signal-to-noise ratio. Therefore, the statistics relating to the PSM scores and information contained within the spectra were investigated. As the target intensity is decreased, the mean PSM score (Figure 46A) and the proportion of fragment ion intensity that can be annotated to theoretical fragment ions (Figure 46) also decrease, indicating lower quality spectra. Because of this reduction in spectra quality, lower values of target intensity were not explored. For further experiments, a target intensity of 10,000 with an intensity threshold of 2,000 were used.

Overall, while these optimisation efforts improved the proteome depth achieved at a throughput of 100 samples per day, the improvements were only modest compared to the proteome depth (and completeness) required for spatially resolved proteomic analysis.

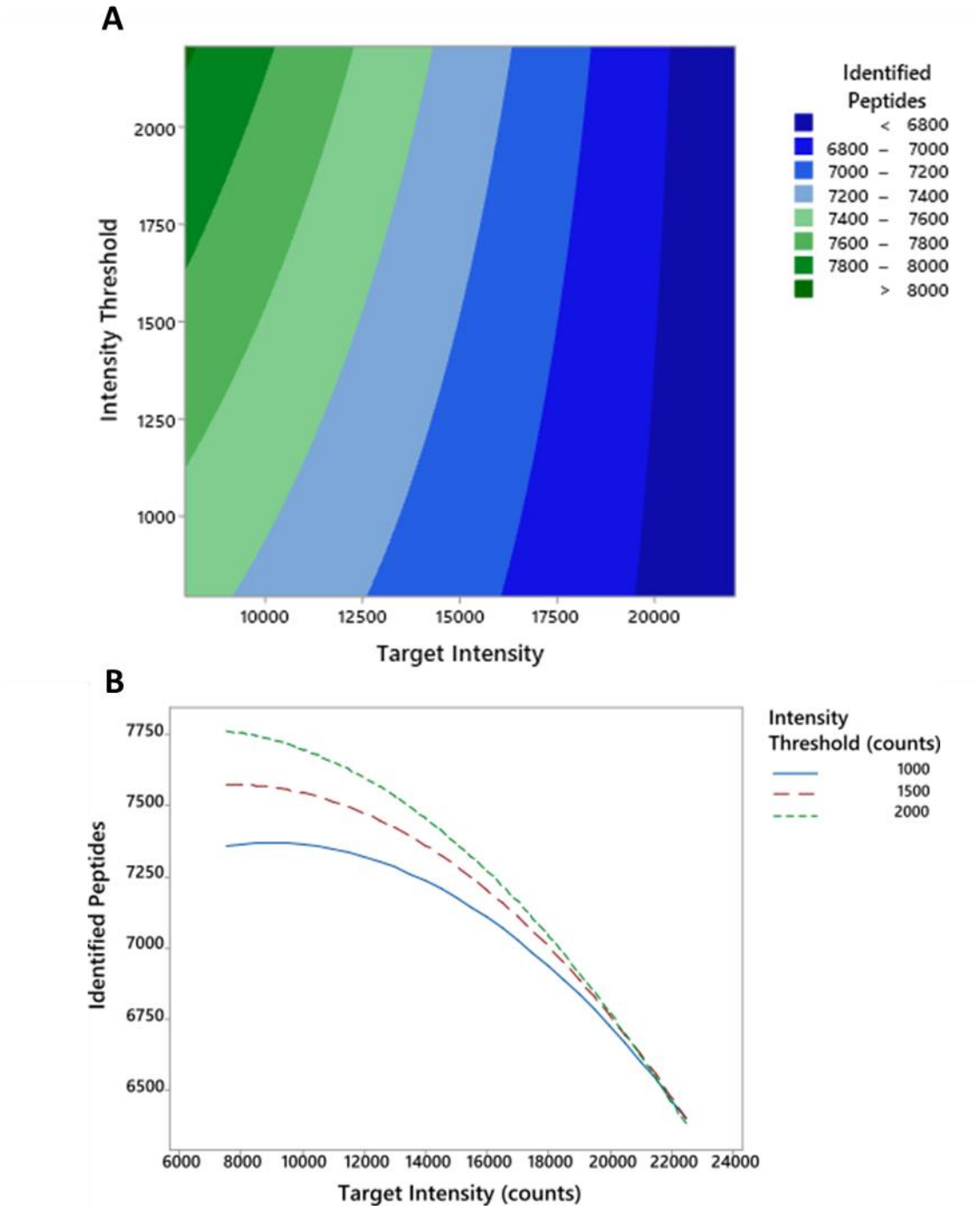

Figure 45 – Optimisation of instrument PASEF MS/MS fragmentation intensity threshold and target

(A) Contour plot visualisation of response surface for the number of identified peptides when varying the target intensity and intensity threshold values relating to triggering MS/MS of precursors. (B) Interaction between target intensity and intensity threshold.

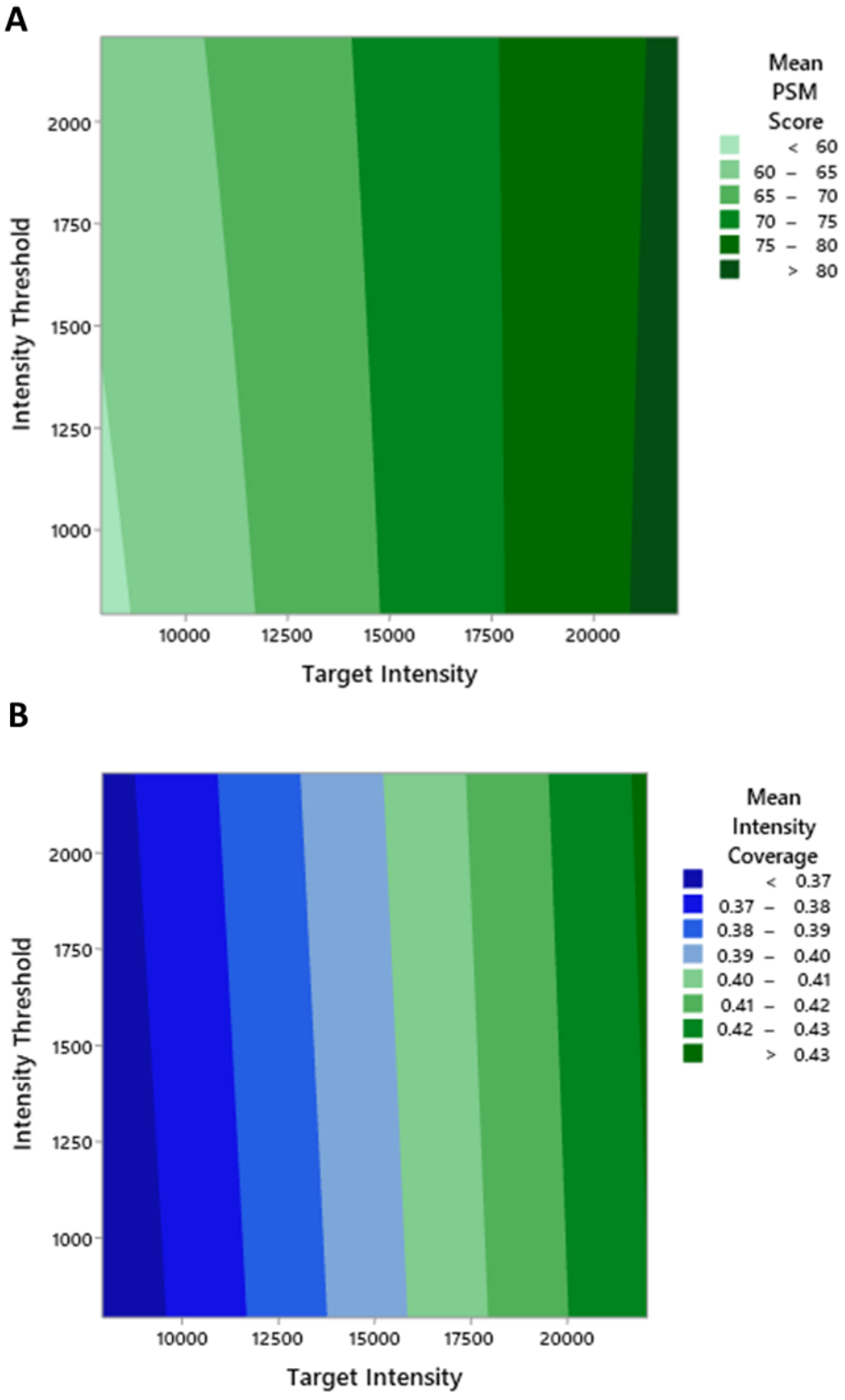

Figure 46 – Effect of PASEF MS/MS fragmentation targets on PSM quality

Contour plot visualisations of response surfaces when varying the target intensity and intensity threshold values relating to triggering MS/MS of precursors. (A) Response surface of mean PSM score. (B) Response surface of mean proportion of fragment ions annotated to theoretical fragment ions.

4.3.3 Short-gradient nLC/MSMS: A middle-ground for depth and throughput?

The EvoSep/TimsTOF configuration lacks proteome coverage compared to the Dionex/Lumos system. The microlitre flow rates of the EvoSep result in a dilution of the analytes at the point of ionisation, and it is known that ESI produces a concentration-dependent response in the mass spectrometer. In addition, the EvoSep uses LC columns with wide inner diameters and larger C18 bead sizes to aid with robustness, which results in broader elution peaks than with narrower columns and smaller beads and a lower response in the mass spectrometer. To determine whether a middle ground between the two systems is possible, a Dionex LC was connected to the TimsTOF using a high-performance commercial nLC column running at 400 nL/min[32]. This was run in a 'direct injection' setup, without using a trap column (the standard setup in our lab uses trap columns for increased robustness) and with a smaller, 5 µL sample loop that is switched out-of-line once the sample has left the loop to reduce gradient delay. To test this setup's performance compared to the Dionex/Lumos, sets of 8 replicate injections of a HeLa peptide standard were analysed on 17-minute gradients on the Dionex/TimsTOF and on 60-minute gradients on the Dionex/Lumos. A mean of 17,584 peptides from 3,925 proteins were identified using the Dionex/TimsTOF and 21,405 peptides from 3,852 proteins were identified using the Dionex/Lumos. The quantitative performance of this setup was also comparable between the systems, the mean Pearson correlation coefficients (PCC) of the Dionex/TimsTOF and Dionex/Lumos protein LFQ values were both 0.99, with median %CV values of 8.0 % and 7.1 % respectively. The Dionex/TimsTOF configuration generates data at a depth equivalent to the Dionex/Lumos with a throughput increase of ~3.7x.

One limitation to this direct injection approach is that an increase in the volume of sample injected will impact the throughput of the sample analysis due to needing to push more sample out of the loop before switching. Increasing the injection volume to 4 µL required 12 minutes of flow for the sample to completely exit the loop before switching, giving a total

cycle time of 38 minutes and throughput of ~38 samples per day – still an increase of almost 3x over the Dionex/Lumos.

To determine the performance of the Dionex/TimsTOF LC-MS/MS system configuration, a titration of ATRT tissue areas collected by LCM was digested with SP3 and analysed on the 17-minute gradient. The number of proteins identified compared to the previous titration experiment (Figure 24) is shown in Figure 47. The Dionex/TimsTOF configuration again essentially matches the Dionex/Lumos configuration in the number of proteins identified from areas above $10^{4.5}$ µm^2 (31,000 µm^2) at triple the sample throughput.

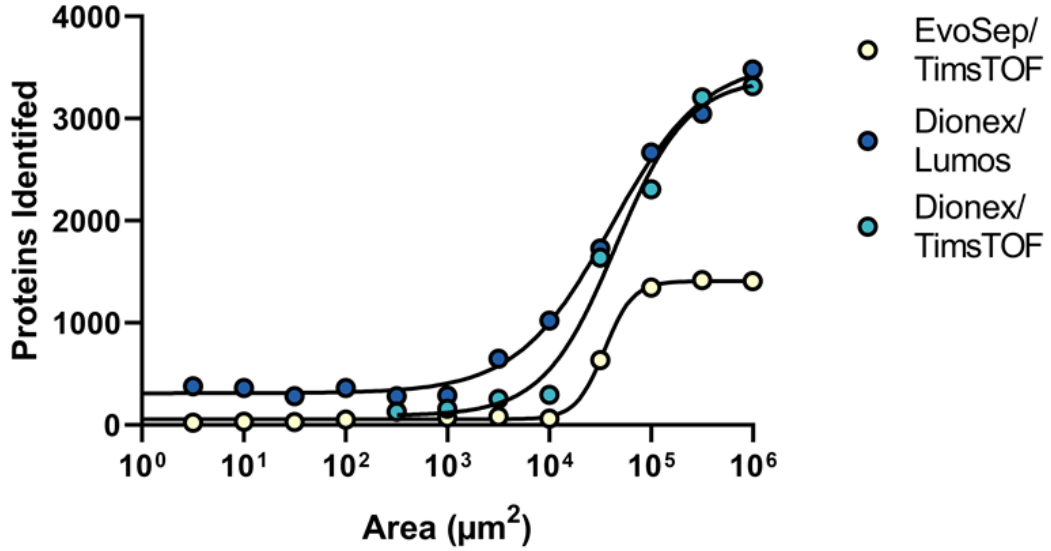

Figure 47 – Protein identifications on three LC-MS/MS systems from a titration of ATRT areas.

Number of proteins identified from a titration of ATRT area on three LC-MS/MS systems. Dionex/Lumos refers to a throughput of ∽13 samples/day on an Orbitrap Fusion Lumos coupled to a Dionex U3000 running at 250 nL/min. EvoSep/TimsTOF refers to ∽100 samples/day on a timsTOF-pro coupled to an EvoSep One running at 2 µL/min. Dionex/TimsTOF refers to ∽38 samples/day on a timsTOF-pro coupled to a Dionex U3000 running at 400 nL/min.

Table 4 – Proteins identified from ATRT tissue titration across LC-MS/MS system configurations

	Area (μm^2)	Side Length (μm)	Evosep/TimsTOF	Dionex/Lumos	Dionex/TimsTOF
$10^{0.5}$	3	2	20	378	-
10^{1}	10	3	30	360	-
$10^{1.5}$	32	6	27	282	-
10^{2}	100	10	51	361	-
$10^{2.5}$	316	18	130	282	127
10^{3}	1,000	32	72	287	158
$10^{3.5}$	3,162	56	82	648	254
10^{4}	10,000	100	63	1020	294
$10^{4.5}$	31,623	178	636	1727	1638
10^{5}	100,000	316	1345	2668	2307
$10^{5.5}$	316,228	562	1419	3047	3208
10^{6}	1,000,000	1000	1407	3480	3318

4.3.4 Spatial proteomics analysis of a large field of view

A section of the ATRT tumour was sampled almost to completion by fitting a 24 x 16 grid (~20 mm x 13 mm) to the tissue to generate 384 samples from 3456 sub-pixels (Figure 48). Each pixel was 833 µm x 833 µm. Samples were collected into 12 cap strips for processing in PCR plates. Samples were analysed using a 17-minute linear gradient using a Dionex U3000 coupled to a timsTOF Pro at a throughput of approximately 40 samples per day. In total, 5,321 proteins were identified in the experiment, with a range of 32 – 4,741 proteins identified per sample. This range includes empty pixels where no tissue was present. 90% of samples with tissue present had over 2,200 proteins identified. Between 35 and 84 proteins were identified in the empty pixels, demonstrating a low level of sample contamination during collection, digestion, and analysis. Each of the 5,321 proteins identified was quantified in at least one sample, with a range of 5 – 3,667 quantified per sample. Again, this is including empty pixels. 90% of samples had over 1,200 proteins quantified. The distributions of identified and quantified proteins per sample are shown in Figure 49A. Figure 49B shows the distribution of quantified values per protein. The data shows a similar distribution to previous experiments

(Figure 41) but with a more proteins with only one quantified pixel. The data completeness of this dataset is 46 %, suggesting that increasing the depth of the proteome coverage does not lead to more complete data while at the depths possible with the acquisition time used here.

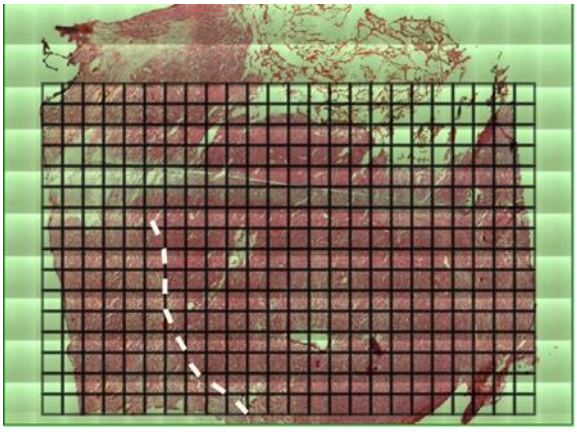

Figure 48 – ATRT slide for large field of view analysis

Composite image of H&E stained section stitched from multiple fields of view acquired on the LCM instrument, with raster sampling in a 24 x 16 grid overlaid. The dashed lines represent the transfer of the pathology annotation to this section.

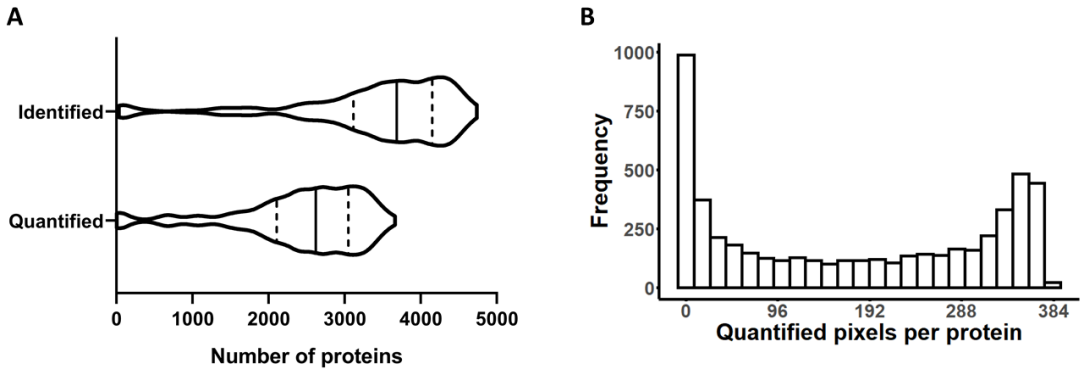

Figure 49 – Distribution of identified and quantified proteins in large region of ATRT

(A) Violin plots showing distributions of identified and quantified proteins per pixel in the large ATRT field of view. Solid vertical lines represent median value. Dashed vertical lines represent lower and upper quartiles. (B) Histogram of quantified pixels per protein map

4.3.4.1 Feature-independent analysis

The value of Moran's *I* was calculated for the 4,306 proteins quantified in at least 9 pixels. Of these proteins, 3,212 proteins have a *q*-value ≤ 0.05, and 2,726 of those proteins have a *p*-value equal to 0.001 (i.e. none of the permuted value of *I* exceed the observed value of *I*). The pixels containing the large region of haemorrhage and some empty pixels with no tissue were excluded from this and further analysis.

As the number of proteins called as significant from this experiment is very large, the distribution of summed and mean sample intensities were also tested for autocorrelation to determine whether the input amount varied across the tissue, which could explain the large number of significant proteins. These summed and mean intensities were $\log_2$ transformed, mapped to their spatial coordinates and Moran's *I* calculated. Neither the summed ($I = 0.003$, $p = 0.317$) or mean ($I = 0.003$, $p = 0.313$) intensities show significant evidence for spatial autocorrelation and these intensity maps are shown in Figure 50 (maps with values overlaid on each pixel are shown in Figure S17). This suggests that the sampling process worked well, with the exception of two pixels, and is not a cause of systematic spatial variability.

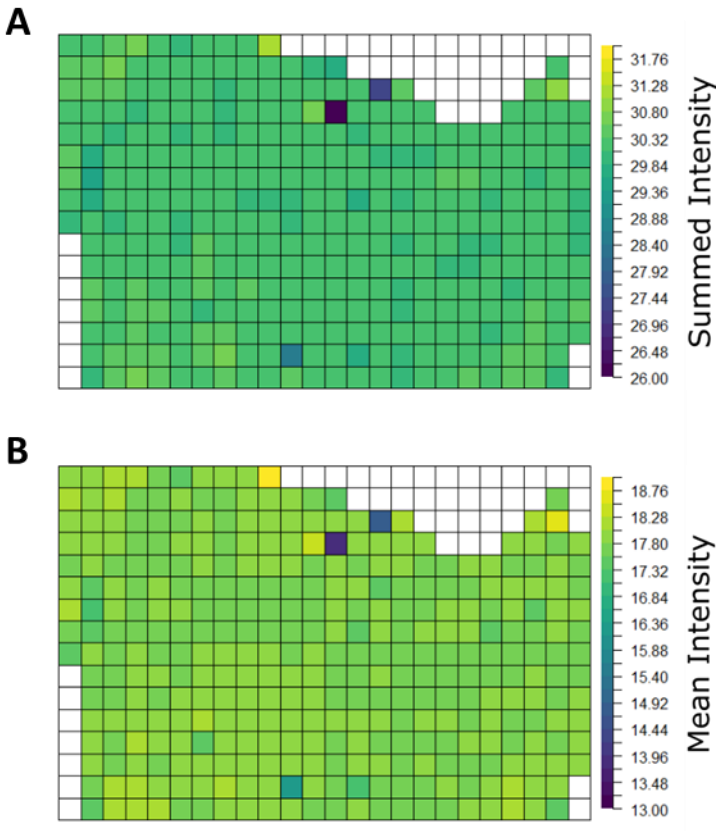

Figure 50 – Summed and mean intensities of each sample

Maps of the $\log_2$ transformed (A) summed and (B) mean intensities of each sample. Samples corresponding to haemorrhage and empty pixels were removed (white).

Returning to the four example proteins of peripherin, PYGL, haemoglobin and histone H4, these proteins have Moran's *I* values of 0.56, 0.58, 0.33 and 0.17, all with *q*-values of 0.002 indicating that all four have significant evidence of spatial autocorrelation. In the maps of peripherin and PYGL, the same pattern viewed in previous experiments can be seen, with the pathology border separating the intense regions of these proteins. In the large field of view, it is the general case that peripherin in and PYGL show opposite expression patterns, with peripherin more abundant near the edge of the field of view and PYGL found with higher abundance in the centre. For haemoglobin, it is most abundant in the region of haemorrhage towards the top right of the field of view. Across the rest of the field of view, its abundance is reasonably consistent with some increases towards the periphery. For histone H4, the region

analysed by previous experiments shows a consistent abundance with similar intensities immediately on either side of the pathology annotation. Moving away from the annotation, the abundance becomes more variable and has a lower intensity in the region of haemorrhage as would be expected.

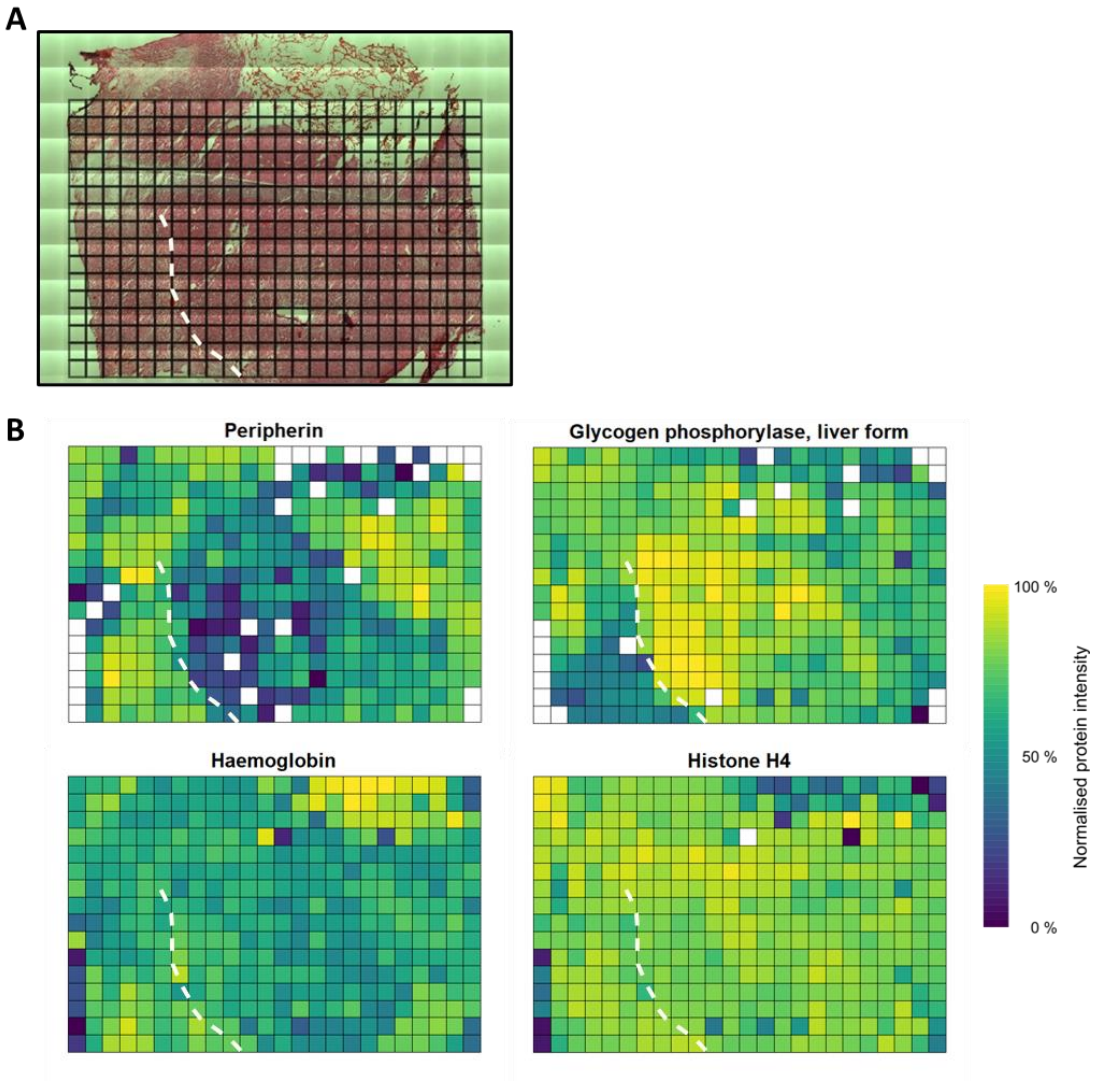

Figure 51 – Spatial proteomic maps of example proteins in a large field of view

(A) Composite image of H&E stained section stitched from multiple fields of view acquired on the LCM instrument, with raster sampling in a 24 x 16 grid overlaid. (B) Normalised protein intensity maps of four example proteins. Normalised protein intensities are scaled separately for each protein. The dashed lines represent the transfer of the pathology annotation to this section.

As the empty and haemorrhage-containing pixels have been removed from the analysis, these are not driving the significant spatial autocorrelation of haemoglobin and histone H4. The

Local Moran's *I* test was used to determine any local clusters of differential expression (Figure 52). Haemoglobin shows several instances of local clusters of high and low values, with a region in the lower-right of the tissue being a cluster of high values surrounded by other high values ('high-high') with a cluster of high values surrounded by low values in the lower right. In addition to these two clusters of high values, there are also two clusters of low values. These clusters could indicate local variations in the amount of vasculature and its permeability. Histone H4 also shows some local clustering of high values, the two pixels in the upper right region correspond to a region that also has an increase in immune cell marker proteins (see & Figure 59 in "4.3.4.3 Clustering-based pathway analysis" below) and these higher intensities of histone H4 could represent a local increase in cell number in this region. The lower number of locally significant pixels in the histone H4 map compared to haemoglobin is consistent with histone H4's lower value of *I* (0.17 vs 0.33).

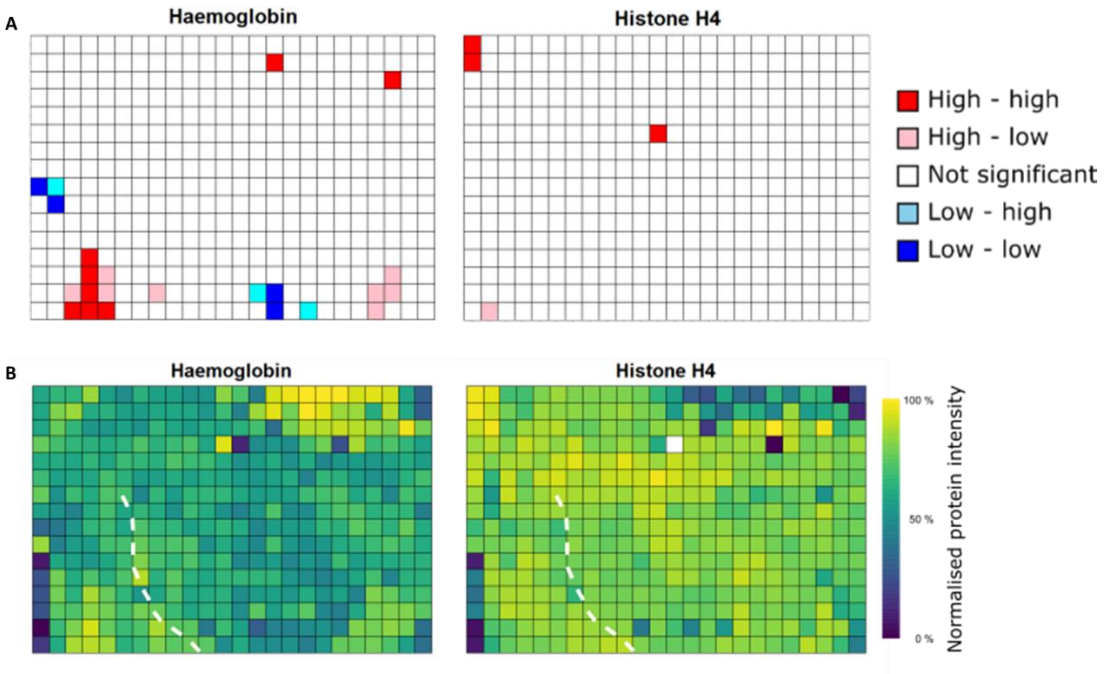

Figure 52 – Local Moran's *I* spatial autocorrelation analysis of haemoglobin and histone H4.

(A) Pixel classification by local Moran's *I* analysis. Pixels are classified based on i) their level relative to the overall mean and ii) the level of their neighbourhood to the overall mean. (B) Normalised protein intensities. Normalised intensities are scaled separately for each protein. The dashed lines represent the transfer of the pathology annotation to this section.

4.3.4.2 Clustering

As so many proteins showed significance by Moran's *I* testing and there does not appear to be a systematic spatial variation in the total abundance of proteins within the tissue, an explanation for this could be that there is enough variability in tissue composition across the region to result in most proteins truly having significant differential spatial expression. The Moran's *I* test for spatial autocorrelation does not indicate where the locations driving this autocorrelation occur, this can be investigated using the local test. However, the local test involves testing each pixel for each protein. In total, after removing missing values, this dataset has 935,844 data points – one per quantified protein per pixel. Running the Local Moran's *I* test on this dataset would involve one test per data point (comparing the data point to the neighbours). This very large number of statistical tests could result in a substantial number of false positives. Because of this, an alternative approach was examined based on hierarchical clustering of the dataset and ANOVA-based significance testing of the clusters generated.

The dendrogram produced by hierarchical clustering of the non-haemorrhage and non-empty pixels was cut at various heights to generate clusters, and these clusters mapped back to their spatial positions Figure 53. While it eventually generates clusters resembling the tissue histology and example protein expression patterns, this method of simply cutting the dendrogram at a fixed height generates many clusters which contain one or two samples. For example, there need to be 26 clusters before the region of brain/tumour interface was represented as its own cluster. Even though many of the clusters are single samples, there are clusters that are contiguous.

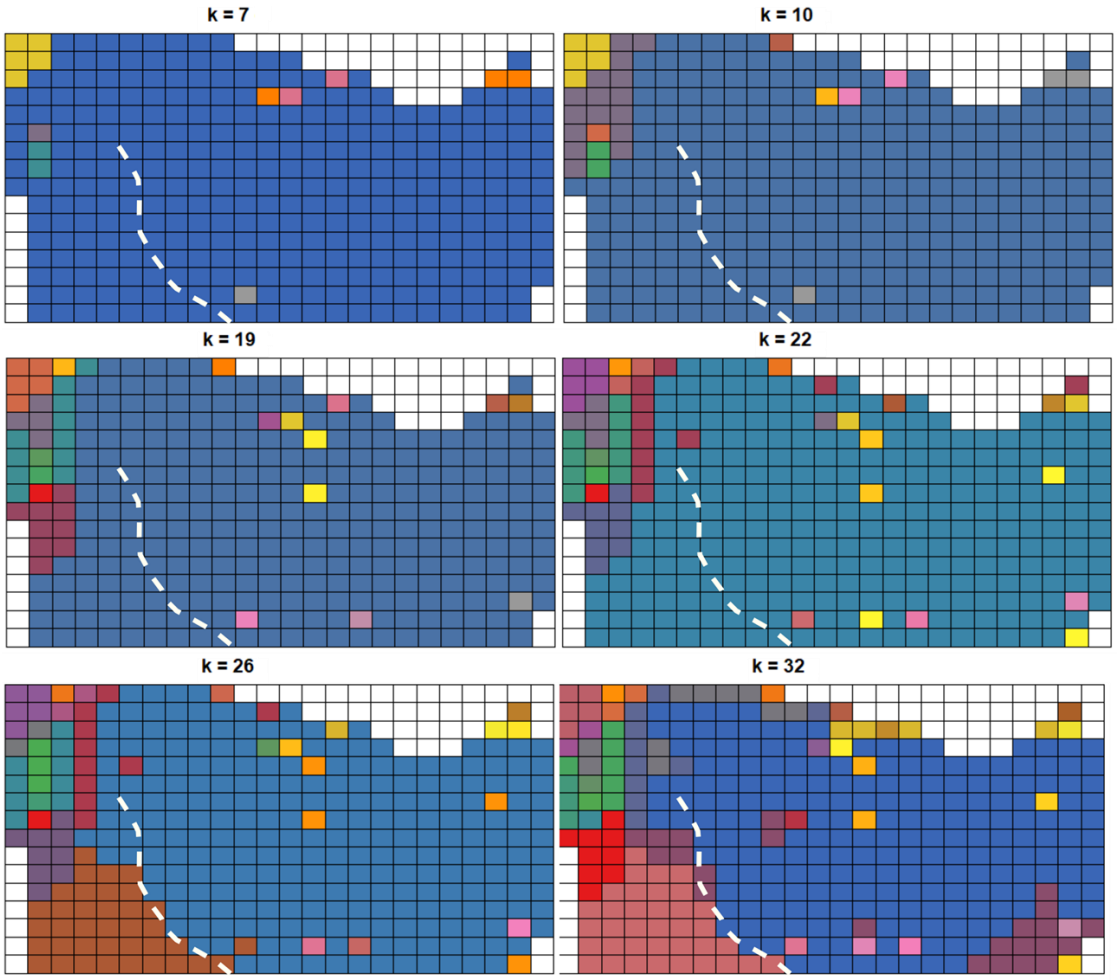

Figure 53 - Spatial clustering of samples

Maps of cluster assigned based on hierarchical clustering of samples at *k* clusters. Cluster membership is shown using colour. White pixels represent removed haemorrhage and empty pixels. The dashed line represents the transfer of the pathologist's annotation to this section

As this clustering approach based on a static cut of the dendrogram showed some promise, a different method of cutting the dendrogram generated by hierarchical clustering was tested. A method that was specifically developed for complex dendrograms is the Dynamic Tree Cut method[257]. This method tries to take the shape of the dendrogram branches into account when assigning clusters. One of the tuneable parameters is the 'height' parameter, representing the allowed distance between members of a cluster – a lower height giving smaller clusters. The results of this clustering method at three different height values are shown in Figure 54. Six prominent clusters are visible at a height of 140, and as this height

parameter decreases to a height of 100, the large cluster is split into smaller subclusters. Unlike the static cut method above, most of the generated clusters at all heights form contiguous regions with no clusters consisting of a single sample. In addition to the cluster maps, the assigned clusters (height = 100) were plotted onto a uniform manifold approximation and projection (UMAP) visualisation (Figure 55A)[263]. The clusters visible in the UMAP visualisation correspond well to the clusters generated by the dynamic cut method (Figure 54, Figure 55B). It should be emphasised that neither the hierarchical clustering, dendrogram cutting and UMAP methods use the sample coordinates. The spatial clustering seen in both methods is driven solely by the quantitative profiles of the samples.

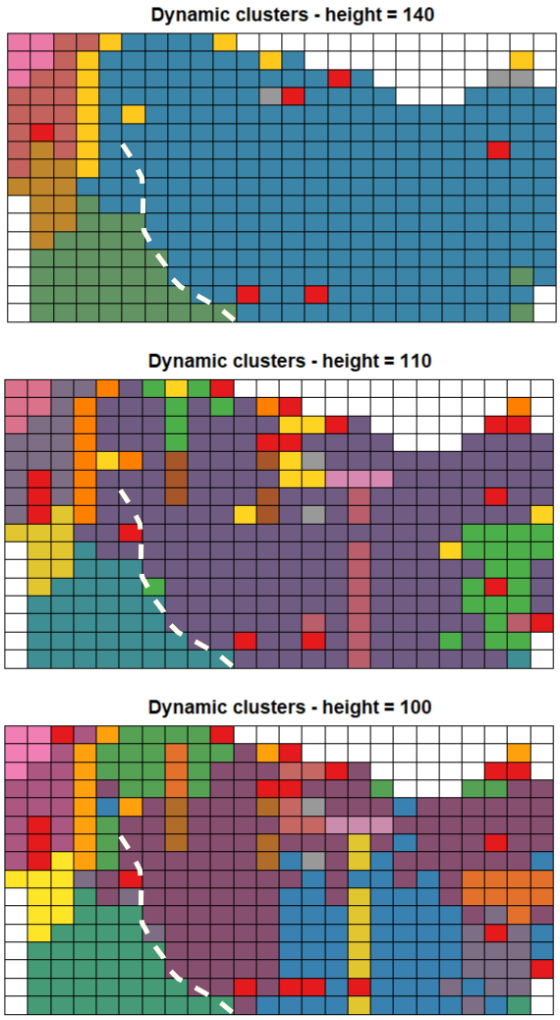

Figure 54 – Spatial clustering of samples using Dynamic Tree Cut

Maps of cluster assignment based on hierarchical clustering of samples with clusters defined at different levels of the height parameter. A lower height generates more clusters. Cluster membership is shown using colour White pixels represent removed haemorrhage and empty pixels. The dashed line represents the transfer of the pathologist's annotation to this section

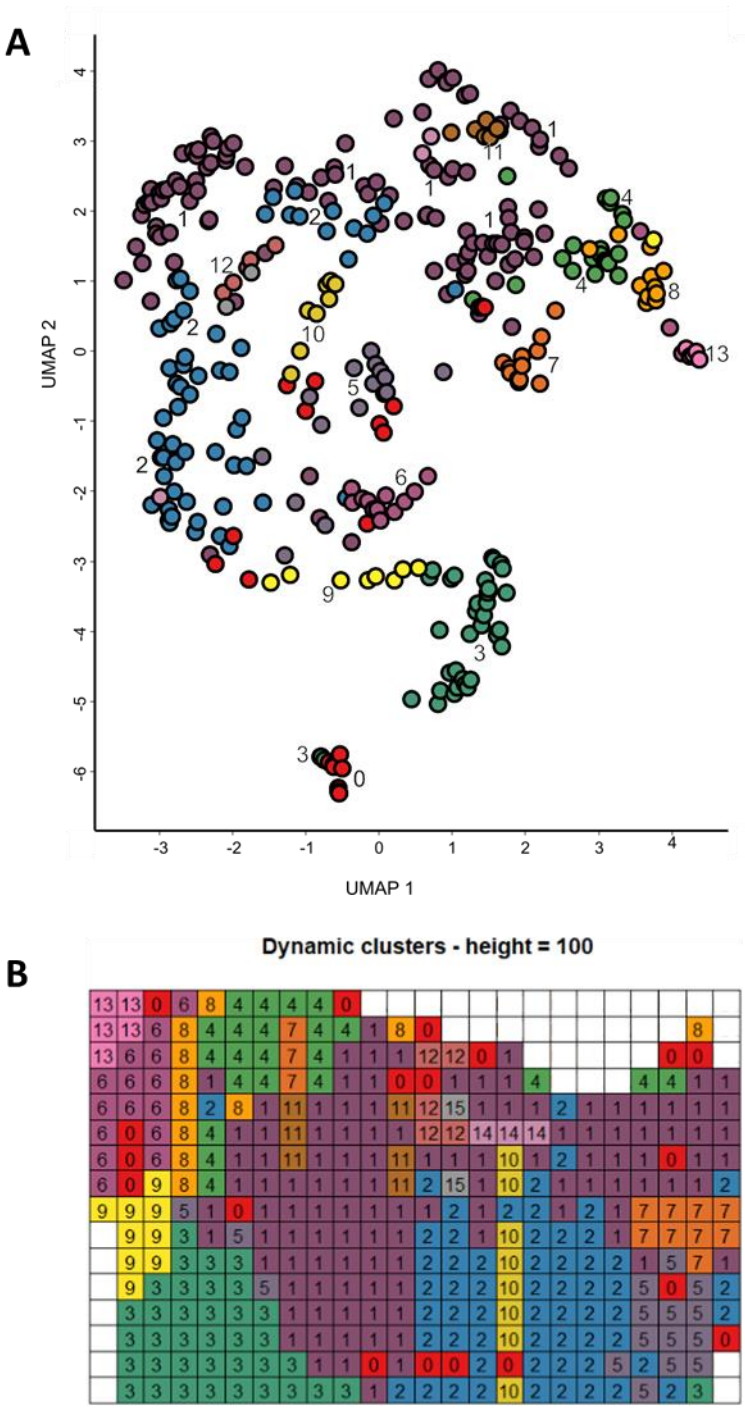

Figure 55 - UMAP embedding of data

(A) UMAP embedding of large field of view data, data points are coloured and labelled according to the clusters generated from the dynamic cluster method at height = 100 (B) Cluster map of dynamic cluster method at height = 100 for reference.

4.3.4.3 Clustering-based pathway analysis

As this clustering method generates spatially well-defined clusters, it allows for a feature-dependent approach without needing prior knowledge – replacing the histopathological features with clusters (features) defined by the spatial proteomic analysis. The clusters generated at height = 100 from the dynamic clustering output were tested for differences using a one-way ANOVA test to determine how protein intensities vary across clusters. Once again, a large proportion (3,512) of proteins have significant evidence for differential abundance between two or more clusters at a q value ≤ 0.05. Pairwise post-hoc testing was performed using the Tukey test for multiple comparisons, resulting in fold-change estimates and associated q-values for each protein in each set of two clusters. These fold-change estimates were used as input to gene set enrichment analysis for the cluster pairs described below to investigate functional differences between the identified clusters.

Pairwise functional analysis of clusters 1, 2, 3, 6 and 13 reveals that processes involved in the immune response, cell adhesion, angiogenesis, development, ECM organisation, mRNA processing and cell cycle vary in abundance across the five spatial clusters. These data are visualised in the form of gene ontology term networks where significantly enriched terms are displayed as nodes in the network, and shared proteins between terms are visualised as edges between the nodes. The comparison between cluster 1 and 3 is shown in Figure 56 and shows that pathways involved in ECM organisation, mRNA processing and translation, cholesterol metabolism and neuronal function are differentially regulated between the clusters. A more global overview is shown in Figure S18. These are generally consistent with the functional analysis of the equivalent regions in the previous chapter (Figure 29).

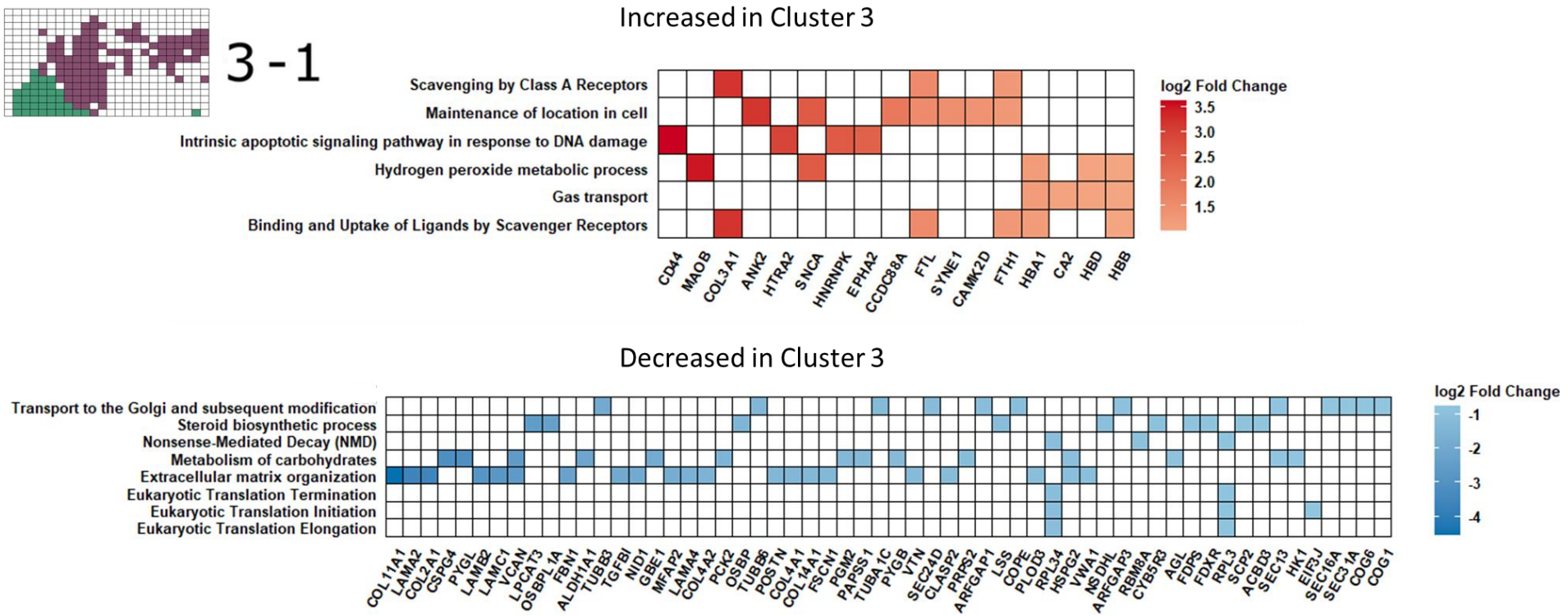

Figure 56 – Functional analysis of Clusters 1 and 3

Gene set enrichment analysis of Gene Ontology Biological process terms with clusters 1 & 3 as defined by hierarchical clustering and dynamic tree cutting. Gene set membership is indicated by colouring the corresponding cell with that gene's $\log_2$ fold change. Log2 Fold Change refers to the estimate of mean difference calculated by a Tukey's *post-hoc* test between the two clusters.

Figure 57 shows the enrichment results between cluster 1 and 2, two large regions of solid tumour (a more global overview is shown in Figure S19). Pathways and proteins involved in ECM organisation, cell adhesion, apoptosis and coagulation are differentially regulated between these two clusters. These differences could represent heterogeneity in the metastatic nature of the tumour cells within these regions, as the presence of blood coagulation enzymes in the tumour microenvironment plays an important role in tumour progression[264]. One of the proteins represented by these coagulation terms is F-actin capping protein subunit beta (CAPZB). CAPZB has been shown to promote tumour progression in the context of epitheloid sarcoma, another SMARCB1-deficient tumour[265].

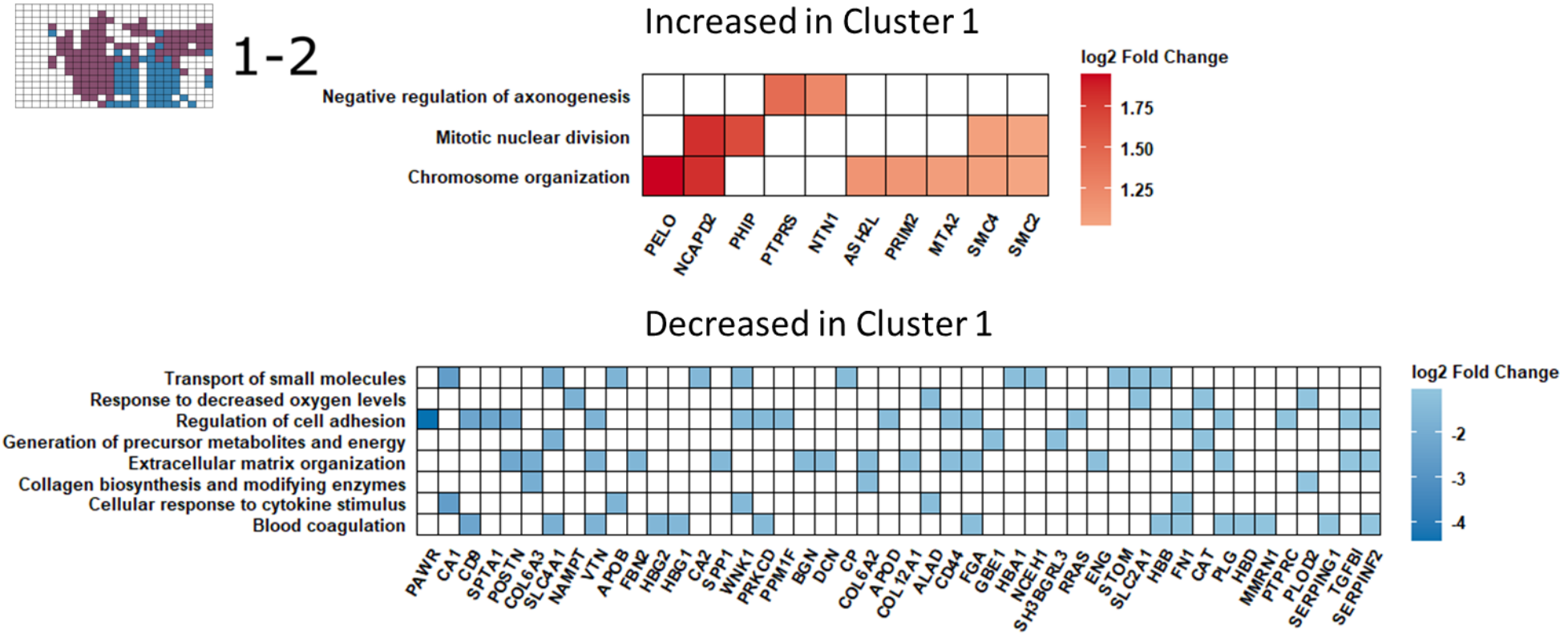

Figure 57 - Functional analysis of Clusters 1 and 2

Gene set enrichment analysis of Gene Ontology Biological process terms with clusters 1 & 2 as defined by hierarchical clustering and dynamic tree cutting. Gene set membership is indicated by colouring the corresponding cell with that gene's $\log_2$ fold change. Log2 Fold Change refers to the estimate of mean difference calculated by a Tukey's *post-hoc* test between the two clusters.

Figure 58 shows a region in the upper left of the tissue section with a large enrichment in immune system pathways and could represent a site of immune cell infiltration (A more global overview is shown in Figure S20). Several immune cell markers are highly abundant in this section of the tissue, for example: neutrophil cytosolic factors 2 and 4, neutrophil elastase and CD45 (Figure 59A). In addition to plotting the intensities of individual proteins that form part of these identified pathways, the pathways themselves can also be visualised in the same manner if some value can be assigned to the pathway for comparison between individual samples. One technique that provides a score for samples against molecular signatures is single sample gene set analysis (ssGSEA)[260]. Essentially, the ssGSEA method takes a ranked list of genes ordered by abundance and analyses the distribution of a set or sets of genes (e.g. a specific pathway, ontology term or cell-type markers) to determine whether they are enriched at the high or low end of the ranked abundance list to generate an enrichment score. As the name suggests, ssGSEA can be performed on every sample in a dataset, allowing the spatial mapping of the enrichment score. Figure 59B shows one such map generated from a gene list of neutrophil marker proteins derived from single-cell RNAseq analysis of bone marrow [266]. This cell-type map shows a spatial distribution similar to the three neutrophil marker proteins in Figure 59A and is a promising proof of concept for an approach to increase the interpretability of spatially resolved proteomic data by visualising pathway-level and cell-type variation within tissue. In glioblastoma, multiple studies have shown that multiple molecular subtypes are present within the same tumour[267–269]. One interesting application of these gene set maps would be to map the marker proteins which correspond to the molecular subtypes of ATRT to investigate whether ATRT tumours are also comprised of cells of multiple subtypes.

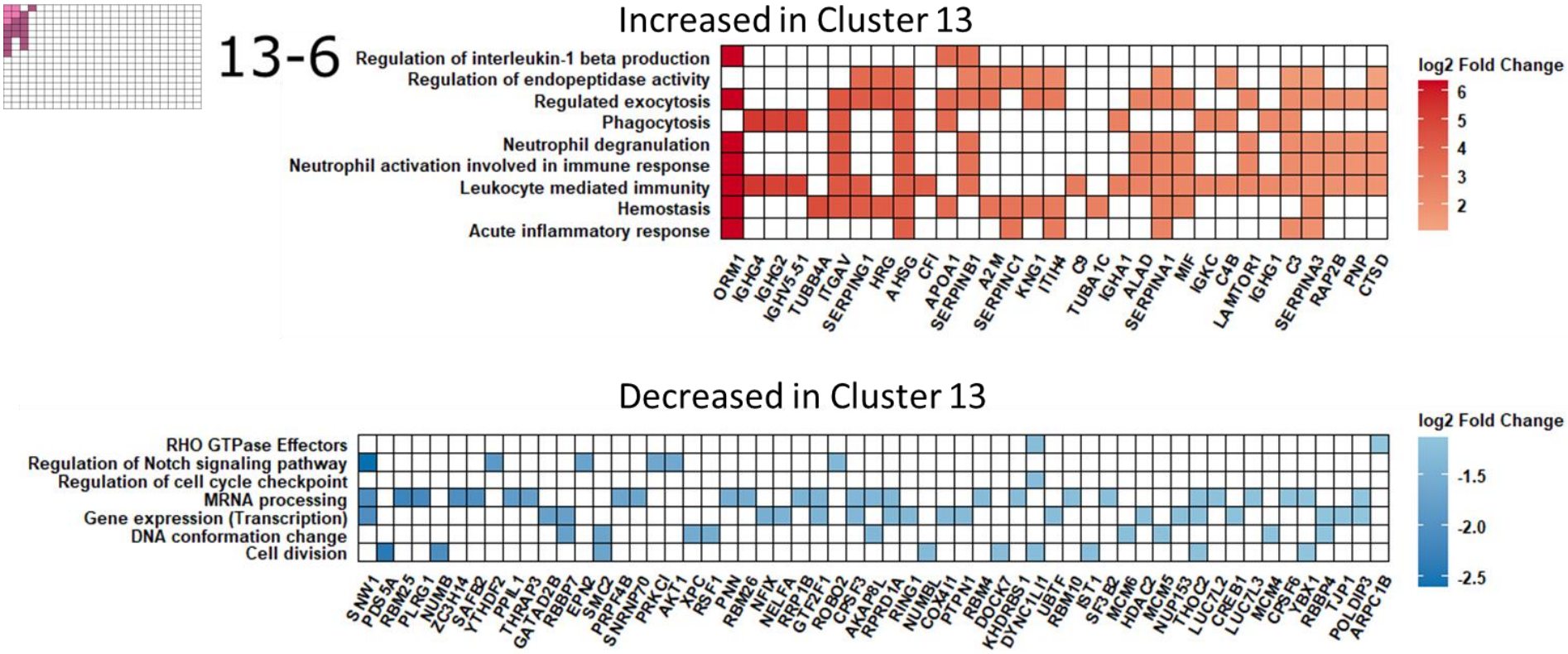

Figure 58 - Functional analysis of Clusters 6 and 13

Gene set enrichment analysis of Gene Ontology Biological process terms with clusters 6 & 13 as defined by hierarchical clustering and dynamic tree cutting. Gene set membership is indicated by colouring the corresponding cell with that gene's $\log_2$ fold change. Log2 Fold Change refers to the estimate of mean difference calculated by a Tukey's *post-hoc* test between the two clusters.

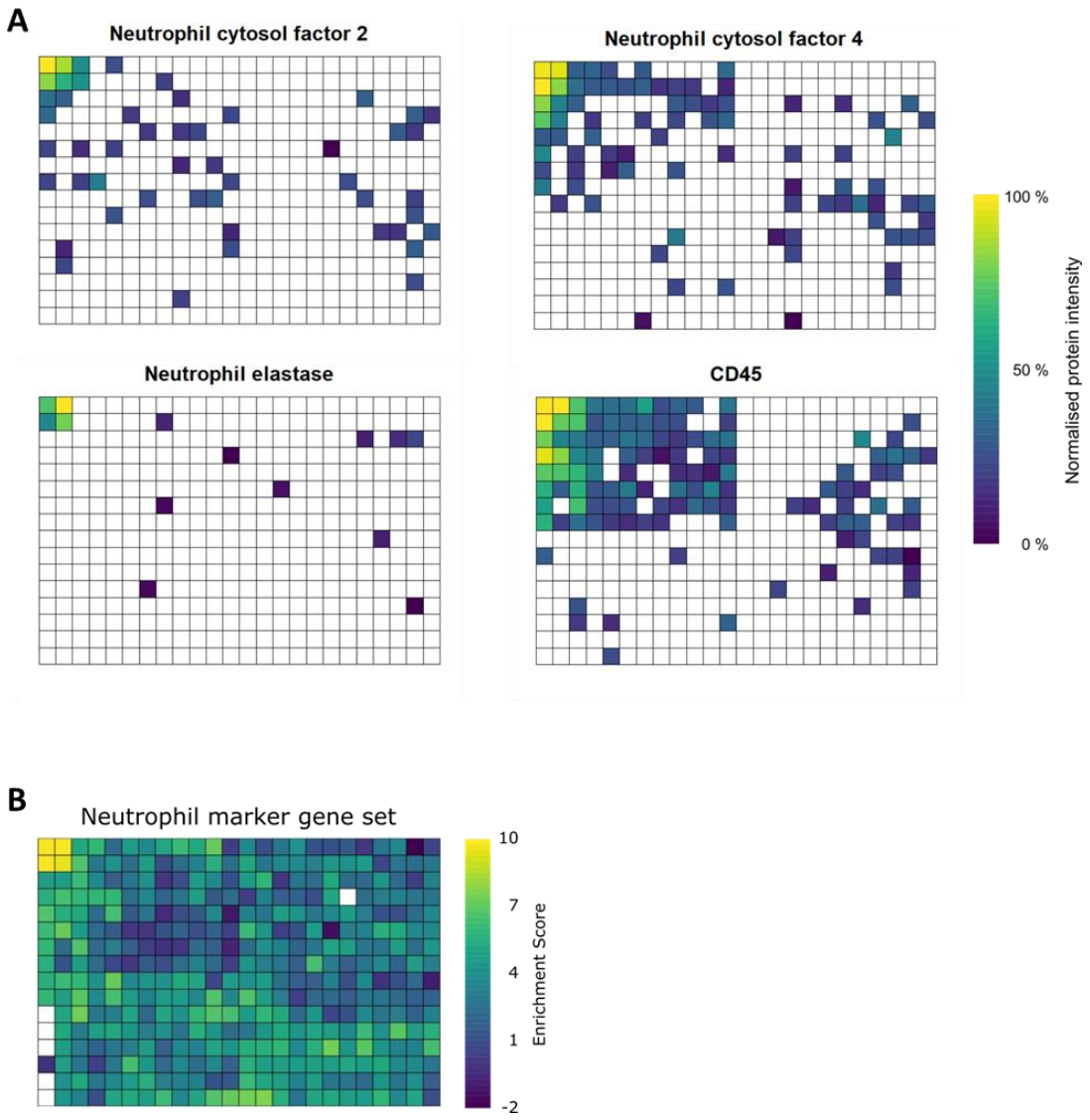

Figure 59 – Spatial proteomic maps of immune cell marker proteins identified from cluster analysis

Spatial proteomic maps of neutrophil and immune cell marker proteins identified as differentially abundant in the hierarchical clustering analysis. Normalised protein intensities are scaled separately for each protein.

Taken together, this clustering method, in combination with subsequent pathway analysis, generates clusters that represent functionally distinct regions of tissue and has good potential for use in the unbiased analysis of spatially resolved proteomic data without the need for prior knowledge of tissue structure and composition.

4.3.5 Posttranslational modifications

In addition to quantifying the expression of peptides and proteins within a sample, mass spectrometry-based proteomics can also be used to investigate posttranslational modification (PTM) statuses of those proteins. One PTM of general interest in cancer biology is trimethylation of histone H3 lysine 27 (H3K27me3). H3K27me3 has been shown to inactivate gene transcription through chromatin compaction and is catalysed by the PRC2 complex[270]. ATRTs have previously been shown to be depleted of H3K27me3 compared to normal brain tissue, and re-expression of SMARCB1 in ATRT cell lines resulted in increased abundance of the H3K27me3 PTM by mass spectrometry [271,272]. None of the PRC2 complex proteins were detected in the spatially resolved proteomic data acquired from ATRT.

In order to investigate whether the data generated could be used to profile histone modification status throughout the region sampled, the intensity distribution of the H3K27me3 modification within the first tissue section was tested for spatial autocorrelation by re-searching the existing mass spectrometry data with the inclusion of a trimethylation modification on lysine (+42.0470 Da). H3K27me3 was not found to be spatially autocorrelated by a Moran's *I* test (I = 0.10, $p = 0.15$). The H3K27me3 modification was only quantified in 32 samples, and many of the quantified pixels have a low number of neighbours, which can hinder the effectiveness of the feature-independent analysis. Therefore, a feature-dependent approach was used to test whether there is a difference in the amount of this modification between the annotated regions of solid tumour and brain/tumour interface. The H3K27me3 modification shows slight enrichment within the solid tumour pixels with a $\log_2$ fold-change of 0.55 ($p = 0.022$). While the levels of H3K27me3 appear to increase in the solid tumour, the level of peptides containing the unmodified K27 site could also differ. In dedicated studies of histone PTM status, the data are usually normalised to the total intensity for all peptides with the same sequence of amino acids, including the modified versions of those peptides. The intensity map for the unmodified H3K27-containing peptide is even more sparse than that of

the H3K27me3 form. This removes the ability to perform such normalisation. Analysing the 10 pixels where the unmodified form was quantified shows an increased expression of the unmodified form of the peptide in the solid tumour (log_2 fold-change = 1.3, two-sample t-test p = 0.025).

The analysis of histone proteins by traditional bottom-up proteomics is challenging because they have a relative enrichment of arginine and lysine amino acids[15]. This enrichment results in tryptic digestion of histones generating many short peptides which are not sequence unique. Methods to overcome this involve the derivatization of lysine to block tryptic digestion at these sites, resulting in longer peptides more amenable to LC-MS analysis[15]. These data suggest that the approach developed here is not suited for the analysis of the histone modification status and that those questions would be better suited with more tailored follow-up experiments. One such experiment could be the larger-scale LCM of identified clusters of expression patterns identified in Figure 54 and sample preparation with lysine derivatisation.

4.3.6 MALDI Imaging of ATRT tissue

4.3.6.1 Peptides

As an attempt to validate the proteomic findings with an alternate technique, MALDI imaging of peptides generated from *in-situ* trypsin digestion of two adjacent ATRT tissue sections was performed. The pixel size was 50 µm and the recorded mass range was 600 – 3,400 m/z. Across the two sections, 498 m/z intervals were detected and taken forward for analysis.

Unsupervised bisecting k-means segmentation was performed to determine clusters of spectra (pixels) showing molecular similarity (Figure 60A). This segmentation was performed on all spectra across both sections and showed reproducibility through corresponding within-section segmentation without between-section segmentation. Segmentation occurs along the margin between the annotated regions of solid tumour (purple regions) and brain/tumour interface (blue regions), reinforcing the presence of molecular differences at the protein level between these two regions. As in the LC-MS/MS raster experiments, the detected molecules

exhibit a range of spatial distributions. Figure 60B shows the spatial distribution of a peak at 1,375.7505 m/z is mainly localised to the annotated region of solid tumour. Figure 60C shows strong localisation of a peak at 931.5480 m/z to the brain/tumour interface. In both of these cases, there is variability of the peak's intensity within the histopathological regions. An example of a peak with a more even spatial distribution is shown in Figure 60D, where the 1205.5668 m/z peak is present at consistent intensity throughout most of the tissue except for some peripheral regions and within the area of haemorrhage. As this data was acquired without MS/MS, determining the source molecules of these ions is challenging.

Additionally, as there is no separation of the analysed molecules sampled by MALDI, the detected peaks could be comprised of multiple ions whose m/z fall within the error range of the measured m/z. The masses of peptides quantified within the whole-slide raster experiment were checked for any which could have generated the 1,375 m/z peak in the MALDI imaging data, resulting in 8 potential peptide sources for this peak (Table 5). The most intense of these peptides is the TTETQVLVASAQK peptide of Histidine-tRNA ligase (HARS). The second most intense is the VVLLTGETSTDLK peptide of the small nuclear ribonucleoprotein U5 subunit 200 (SNRNP200). These two proteins' maps are shown in Figure 61. In general, the spatial distribution appears to correlate well with the ion intensity map in Figure 60B. This highlights, as others have shown, how complementary LC-MS data can be used to validate and provide identification assignment MALDI imaging data [127,273]. In addition, this concordance between both modalities provides further evidence of the validity of the spatially resolved proteomic approach presented here. It would be interesting to target several peptides of interest from the spatial LC-MS data for MALDI imaging with fragmentation (for sequence information) on an instrument with ion mobility capability to validate peptide and protein distributions further while acquiring higher resolution images.

Table 5 – Peptides quantified in whole-slide raster with masses corresponding to 1,375 m/z peak in peptide MALDI imaging data

Mass	Intensity	Spectral Counts	Score	Sequence	Gene	Protein
1374.74	28162000	324	224.09	TTETQVLVASAQK	HARS	Histidine--tRNA ligase, cytoplasmic
1374.76	23760000	329	145.75	VVLLTGETSTDLK	SNRNP200	U5 small nuclear ribonucleoprotein 200 kDa helicase
1374.75	9445000	26	60	AVPREELFVTSK	AKR1A1	Alcohol dehydrogenase [NADP(+)]
1374.73	8564200	5	36.514	FVSPLSWDLVGR	ABCA1	ATP-binding cassette sub-family A member 1
1374.75	6904800	40	154.94	SVVSGSINTVLGSR	PLIN2	Perilipin-2
1374.77	3335000	122	97.277	FLFNLPLNIER	EXOC2	Exocyst complex component 2
1374.75	976730	10	108.24	VVGAQSLKEMVSK	PGD	6-phosphogluconate dehydrogenase, decarboxylating
1374.73	813390	1	72.922	HEGYLGAIGAFLK	PANK4	Pantothenate kinase 4
1374.74	765060	16	94.804	VTLTSEEEARLK	LDHA	L-lactate dehydrogenase A chain
1374.75	480480	2	30.349	QNIGVATFIGLDK	SMC4	Structural maintenance of chromosomes protein 4

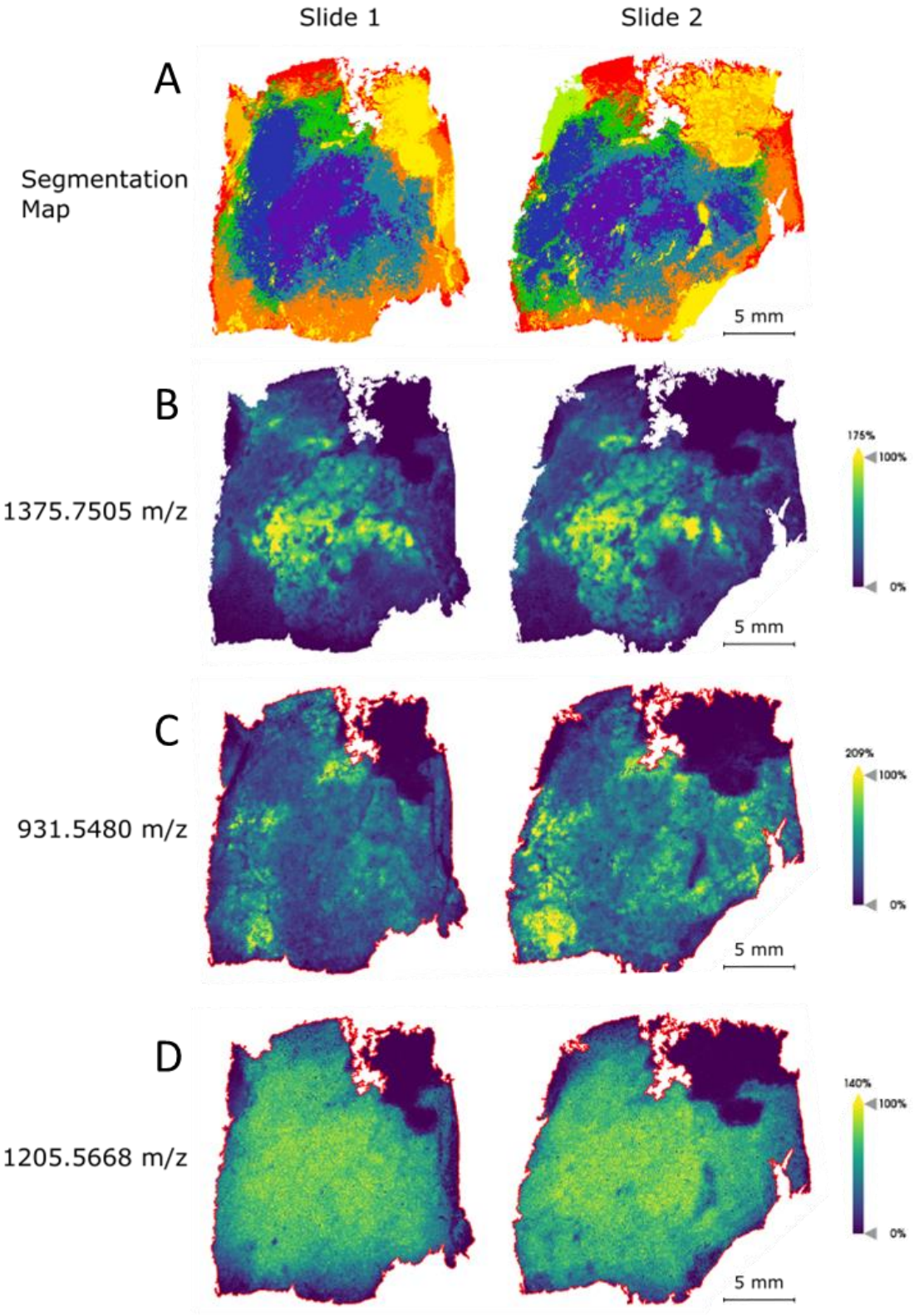

Figure 60 – MALDI imaging of ATRT peptides

Two adjacent ATRT tissue sections were sprayed with trypsin to digest proteins *in-situ* and then imaged by MALDI imaging. (A) Bisecting K-means segmentation of 498 m/z intervals. Segmentation maps were split to resemble the ATRT tissue block histopathology. (B) Ion image of a 20 ppm window centred at 1375.7505 m/z. (C) Ion image of a 20 ppm window centred at 931.5480 m/z. (D) Ion image of a 20 ppm window centred at 1205.5668 m/z.

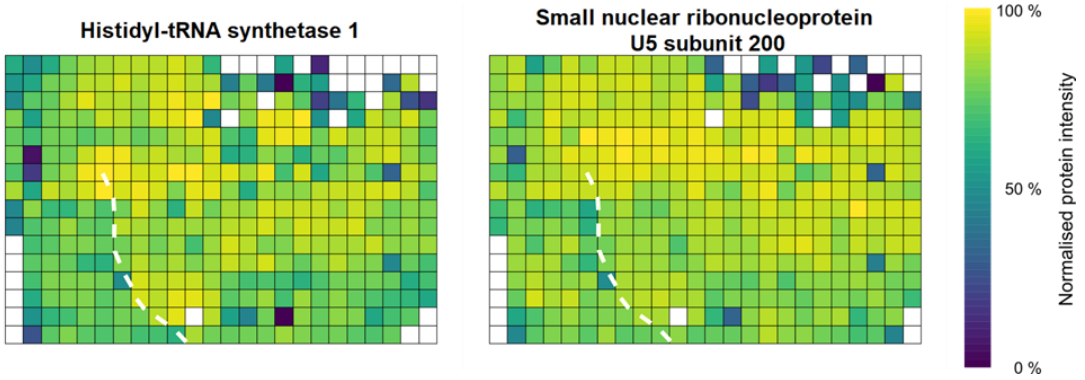

Figure 61 – Candidate proteins for 1,375.75 m/z ion observed in MALDI imaging

Normalised protein intensity maps of proteins with a peptide ion that falls within 20 ppm of 1,375.75 m/z . Normalised intensities are scaled separately for each protein. The dashed lines represent the transfer of the pathology annotation to this section

4.3.6.2 Lipids

In addition to the peptide imaging, lipids were also imaged on a serial section of the ATRT tumour. There were 302 peaks detected across the section, and these were used to calculate an unbiased segmentation map (Figure 62B). The segmentation map was split to resemble the ATRT pathology; the cluster tree is shown in Figure S22. Again, the margin between the two annotated areas is clearly visible. The 616.174 m/z ion of haeme visibly correlated with larger blood vessels visible in the reference image but also highlights the presence of smaller vessels distributed throughout the tissue.

On the macro-scale, the region of solid tumour appears more homogenous in the segmentation map in comparison to the peptide segmentation maps. However, some smaller-scale variation occurs throughout the region. Several ions co-localise with the dominant segmentation mask present in the solid tumour (Figure 62 - C, D, E). Among these, three ions could be assigned putative lipid classes: 852.5832 m/z – PC(38:2); 826.5655 m/z – PC(36:1); and 824.5520 m/z – PC(36:2); PC(36:1) and PC(38:2) have heterogeneous intensity within the solid tumour with a more clustered distribution whereas PC(36:2) shows a most consistent intensity distribution. Several ions co-localise with the brain/tumour interface region surrounding the solid tumour region (Figure 62 – F, G, H). For example: 832.5779 m/z (PC

38:4); 804.5471 (PC 38:7); and 780.5478 m/z (PC 36:5). The interpretation of these is hindered by the lack of an exact molecular identity for these ions, which correspond to groups of molecules within the phosphatidylcholine lipid class. However, many previous studies have shown reprogramming of lipid metabolism in tumour tissue, even going so far as to use this to differentiate tumour tissue from normal tissue in surgery [274,275]. Again, these results show that the ATRT tissue block analysed throughout this thesis is heterogeneous on the molecular level both in ways that are captured by the pathology annotation but also within these annotated regions.

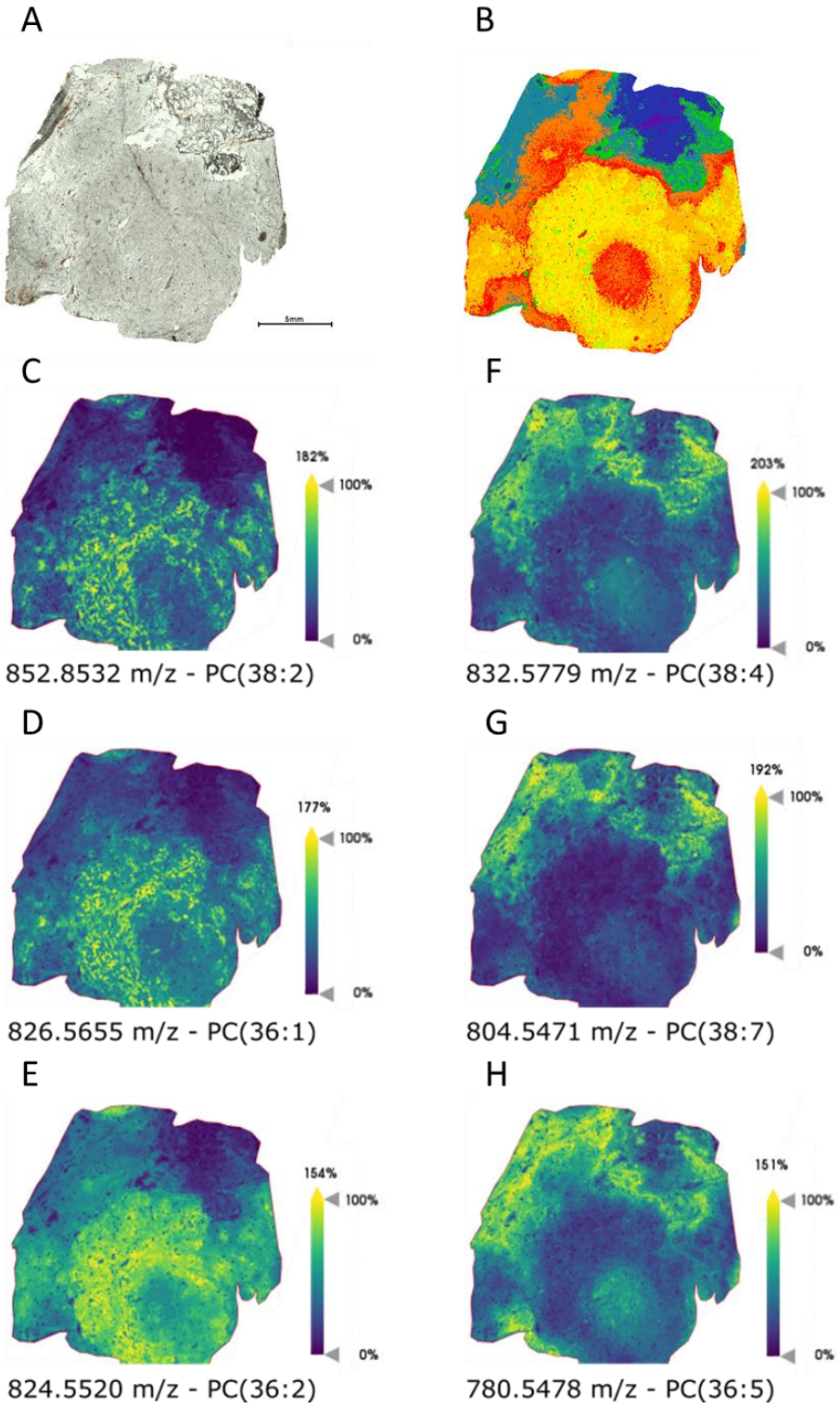

Figure 62 – MALDI Imaging of ATRT lipids

Distribution of several m/z windows putatively assigned to phosphatidylcholine lipid classes. (A) Brightfield image of unstained ATRT tissue prior to MALDI imaging. (B) Bisecting K-means segmentation of 498 m/z intervals. Segmentation maps were split to resemble the ATRT tissue block histopathology. Ion images putatively assigned to: (C) PC(38:2); (D) PC(36:1); (E) PC(36:2); (F) PC(38:4); (G) PC(38:7); (H) PC(36:5).

4.3.7 Validation of spatial proteomic hits by immunohistochemistry

To further validate the spatially resolved proteomics data generated in this work, several candidate proteins showing spatial variation with the maps generated were selected for follow-up immunohistochemistry staining. This was performed in sections taken from the same tissue block used for all the ATRT data presented. Two proteins showing a strong differential expression pattern across their respective maps were chosen: the liver isoform of glycogen phosphorylase (PYGL) (Figure 63 A, B) and aspartate beta-hydroxylase (ASPH) (Figure 63 C, D). ASPH has been shown to be overexpressed in tumours, including other forms of brain tumours, and is involved in promoting cell mobility via the Notch pathway[276,277]. In addition to these two proteins, CD45, a marker protein for immune cells was also stained for (Figure 63 E, F).

The IHC staining closely reflects the spatial protein intensity distributions within the proteomic maps for these three proteins. Both PYGL and ASPH IHC images show strong staining in the region of solid tumour (to the right of the annotation line), with weaker staining in the region of brain/tumour interface to the left of the annotation line (higher magnification of brain/tumour interface is shown in Figure S23). More of the brain/tumour interface cells show PYGL staining compared to ASPH, which only shows staining in a few scattered cells. This reflects the larger difference between solid tumour and brain/tumour interface in ASPH (log_2 fold-change of 4.0, $q < 0.001$) in comparison to PYGL (log_2 fold-change of 3.1, $q < 0.001$) in the comparison of spatial clusters 1 and 3. CD45 staining also closely reflects its intensity distribution in the spatially resolved proteomic data, with strong staining in the upper-left region of the tissue.

These examples show a high degree of concordance with the spatially resolved proteomic data, but further work could increase confidence. For example, staining for more highly abundant proteins within the brain/tumour interface – this work is ongoing, but further antibody optimisation is required. It would also be interesting to stain other cell markers such

as endothelial cells to visualise the distribution of vasculature within the tissue and correlate this with the proteomic data.

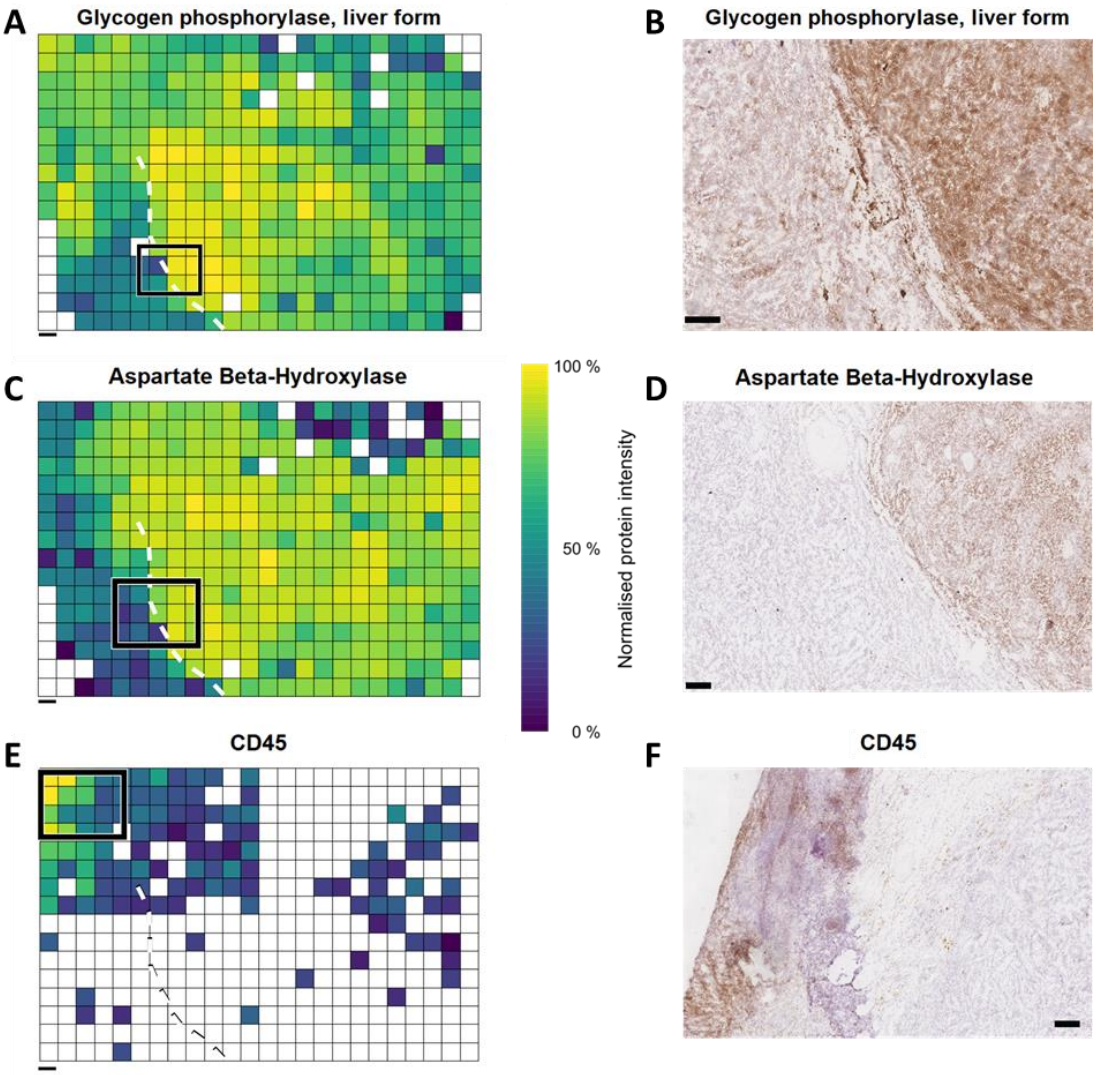

Figure 63 – Immunohistochemistry validation of spatial proteomic maps

Spatial proteomic maps of proteins targeted for follow-up IHC staining. (A,C,E) Normalised protein intensity maps. Normalised intensities are scaled separately for each protein. Rectangle represents the approximate area of the adjacent section shown in the corresponding IHC images. The dashed lines represent the transfer of the pathology annotation to this section, with the brain/tumour interface to the left and the solid tumour to the right. Scale bar = 833 µm. (B,D,F) ATRT tissue stained and visualized by IHC. Scale bar represents 200 µm (B) or 250 µm (D,F).

4.3.8 Reanalysis of public data

In order to determine the applicability of the data analysis workflows that I have developed in this thesis, MaxQuant outputs from Piehowski *et al.* were downloaded and analysed with the spatial proteomics tools [139]. The Piehowski dataset of 2,298 protein maps was generated by raster sampling tissue from the uterus of a pregnant (day 4) Wnt5a-null mouse at 100 µm resolution in a 4 x 6 grid. Samples were digested using their nano well-based nanoPOTS method. Figure 64A shows an overview of the tissue section and sampled region false-coloured based on the dominant cell type of that pixel. Using these dominant cell-type annotations to determine differentially expressed proteins in a feature-dependent manner results in 339 significantly differentially expressed between stromal cells and the luminal epithelium. Of these 339 proteins, 68 have a $\log_2$ fold-change of at least ± 1. When applying the feature-independent method, 224 proteins ($q \leq 0.05$) show significant evidence of spatial autocorrelation and 121 proteins have a *p*-value for the Moran's *I* test of 0.001, indicating that none of their permutations were more correlated than the observed data.

Applying the clustering method described above to the Piehowski data results in three clusters (Figure 64B). These clusters capture two features of this mouse uterus tissue: a dominant region of the luminal epithelium (LE) (Blue, cluster 3) within the stroma, and the mesometrial-antimesometrial axis (red – green, cluster 1 – 2). Figure 64C shows the protein maps for four example proteins, the results of the feature-dependent and feature-independent tests are presented in Table 6. Arachidonate 15-lipoxygenase and coronin 1a are two proteins that are differentially expressed between the LE and stroma. Albumin, along with other blood and plasma proteins, shows a higher level of protein intensity at the bottom of its expression map, corresponding to the antimesometrial pole. This is consistent with the known timeline of mouse pregnancy, as, at day 4, stromal cells begin decidualisation which triggers increased vessel permeability and neovascularisation at the antimesometrial pole[278,279]. The

heterogeneous nuclear ribonucleoprotein A3, and other RNA-binding proteins, show increased expression at the mesometrial pole in a reversed gradient compared to albumin.

Applying the feature-independent method to this previously published dataset further demonstrates its utility to detect proteins with spatially differential expression driven by underlying tissue features without the need for prior knowledge.

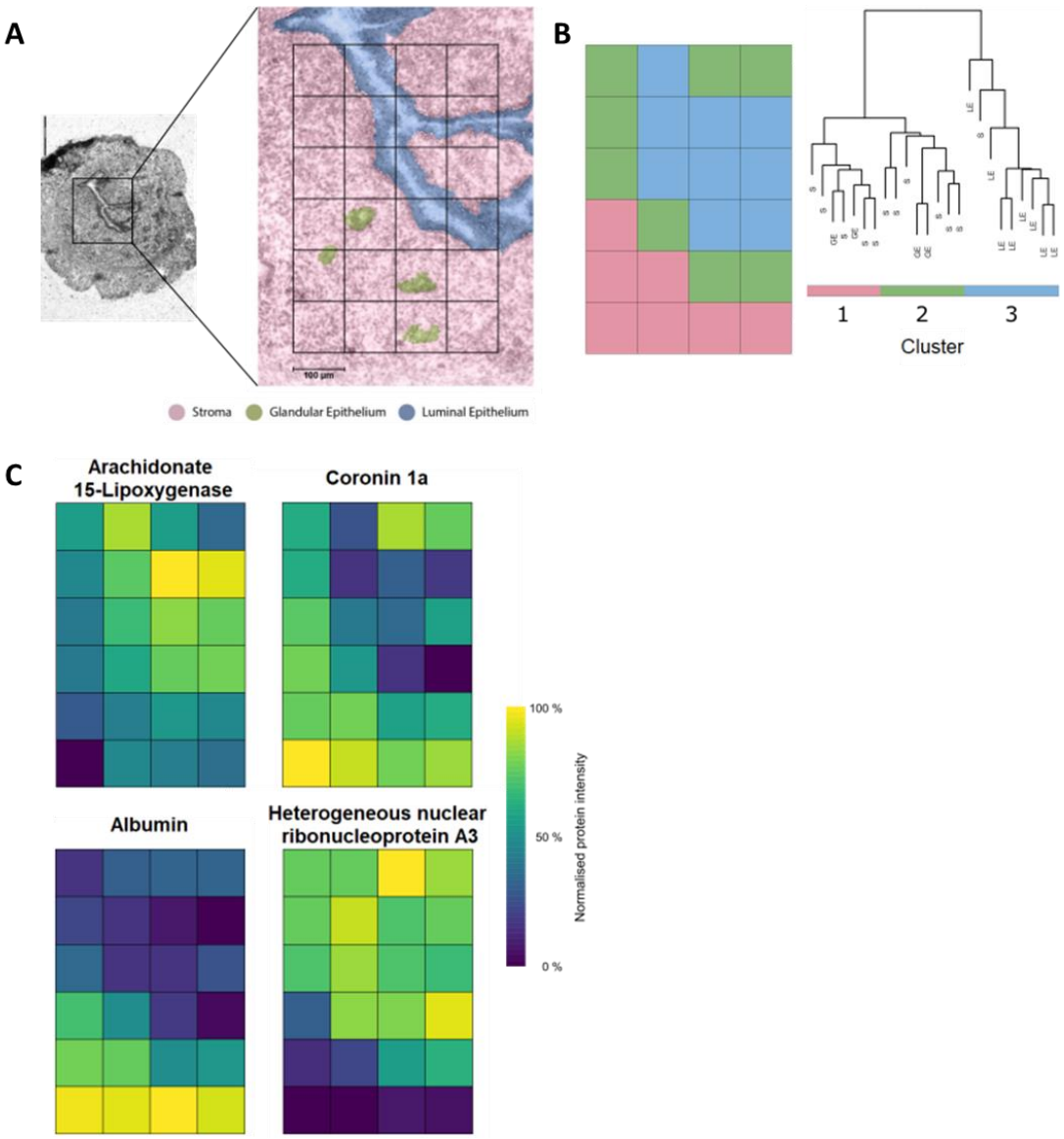

Figure 64 – Spatial proteomic analysis of public mouse uterus data.

Reanalysis of mouse uterus spatial proteomic data from Piehowski *et al.* (A) Overview of Wnt5a-null mouse uterus section (left) with region which was raster sampled (right). Image is false-coloured based on cell type. (B) Cluster map generated from spatial proteomic data (left) and the associated dendrogram (right). (C) Normalised protein intensity maps of four example proteins. Normalised protein intensities are scaled separately for each protein.
S – Stroma; GE – Glandular epithelium; LE – Luminal epithelium.
Adapted from[139]

Table 6 – Summary of four example proteins from Piehowski *et al.* mouse uterus data

Protein	Feature-dependent	Feature-independent
Arachidonate 15-lipoxygenase	$\log_2$ fold-change = -1.4 $q < 0.001$	Moran's I = 0.39 q = 0.007
Coronin 1a	$\log_2$ fold-change = 1.0 $q < 0.001$	Moran's I = 0.25 q = 0.041
Albumin	$\log_2$ fold-change = 0.46 q = 0.007	Moran's I = 0.61 q = 0.007
Heterogeneous nuclear ribonucleoprotein A3	$\log_2$ fold-change = -0.46 q = 0.06	Moran's I = 0.53 q = 0.007

4.4 Conclusion

In this chapter, I have demonstrated improvements to the spatially resolved, mass spectrometry-based proteomics workflow. It can be used in a high-throughput fashion without compromising the depth of proteomes generated using optimised nano-liquid chromatography methods and a sensitive, fast-scanning mass spectrometer. Furthermore, the method can be used to spatially profile large sections of tissue to identify spatially variable proteins and pathways through the use of unsupervised statistical methods that do not require prior knowledge of tissue composition or pathology. The spatial and molecular patterns exhibited within the ATRT tumour tissue were validated by orthogonal mass spectrometry imaging and immunohistochemistry.

5 Summary and Future Perspectives

This thesis has presented the need to generate highly multiplexed, quantitative, spatially resolved measurements of proteins within a tissue. In order to address this need, I have developed methods for sample preparation, liquid chromatography-mass spectrometry data acquisition and statistical analysis of samples isolated from tissue using laser capture microdissection (LCM).

The methods were demonstrated to perform well at resolutions between ~300 – 800 µm, with the potential for higher resolution with some compromise of proteome depth. Analysis of the spatially resolved data highlights proteins, pathways and cell types showing spatially variable abundance within the context of a rare paediatric tumour - atypical teratoid/rhabdoid tumour – and its surrounding normal tissue. This is all performed without prior knowledge of the tissue composition. The identified heterogeneity within the tumour was validated using orthogonal mass spectrometry imaging and immunohistochemistry. The method is also applicable to other studies where samples have been isolated in a spatially resolved manner and analysed by mass spectrometry-based proteomics. The method performs excellently in comparison to general high sensitivity proteomic workflows for tissue analysis and other spatial proteomic methods performed in this resolution range.

In addition, the sample preparation method is applicable to cell-type resolved proteomics of the human brain. The proteomes of two motor neurons were characterised: the well-studied Purkinje cell and the Betz cell. Several well defined Purkinje cell markers were identified, along with other proteins that these neurons are known to express. Follow-up work in our laboratory has shown that several of the proteins identified in Betz cells were specific to these cells within their local context of the primary motor cortex by immunohistochemistry (Connor Scott, unpublished data). Overall, the work presented here informs on methods for spatially

resolved proteomics of tissue, laying the groundwork for further investigation of tissue proteomes, and highlighting the potential utility of such approaches.

While the approaches used in this thesis can generate proteomic data of good depth from a few hundred cells, these data still have some characteristics of bulk data. The molecular profiles of the pixels are still average profiles of the cells contained within the pixel. Methods for the deconvolution of pixel-level data could help alleviate these averaging events and allow for the profiling of cell composition within the pixels. Recently, techniques have been developed to deconvolute spatial transcriptomics data in combination with single-cell transcriptomic data[280–282]. It would be interesting to apply these or similar techniques to spatial proteomics data. Most molecular signatures sets are generated from transcriptomic data, which does not correlate well with proteomic data (at least at the bulk level)[283]. Perhaps these deconvolution approaches would be better suited using molecular signatures derived from proteomic data of bulk-sorted cells, laser capture microdissected cells of a single phenotype or using one of a range of novel single-cell proteomic methods[83,250,284–286].

One disadvantage of using the feature-independent method of determining the spatial autocorrelation of proteins is that it can be confounded by missing values within the data. It is linked to a general limitation of mass spectrometry-based proteomics, which is that if a protein is not quantified within a sample, it does not mean that protein is not present. The calculation of Moran's *I* does not take missing values into account when measuring spatial autocorrelation, but the pattern of missing and observed values can together form spatial patterns. For example, if a protein is truly present/absent in two substructures, this protein may not be part of the set of significant candidates but would likely be of interest. One way to address this would be to investigate the sets of cluster-unique proteins generated from hierarchical clustering and dimensionality reduction data. A second potential method for

addressing this would be clustering and data imputation methods capable of taking samples' spatial relationships into account [256,287].

As only a single tumour sample was analysed in this thesis, it is hard to generalise the biological insights to ATRT. Therefore, it would be interesting to repeat the analysis on other ATRT cases. At a throughput of ~40 samples/day, analysing ten 12x8 grids would require approximately 25 days of instrument time, and so efforts to increase the throughput will be valuable. Recently, it has been shown that data-independent analysis (DIA) outperforms the data-dependent acquisition (DDA) approach used here, with 3,000 - 5,000 proteins identified from 200 ng of HeLa peptides at a throughput of 200 samples per day [8,288]. If this performance holds with the sample sizes used here, this will represent a 5x increase in throughput. In addition, DIA generally has fewer missing values than DDA, which could help alleviate the issues discussed above.

Integrating multiple data modalities and patient data is critical to the comprehensive understanding of disease processes, and the correct integration relies on powerful bioinformatic tools [289,290]. The complete integration of spatially resolved proteomic data with spatially resolved metabolomic, lipidomic and transcriptomic data, along with high-resolution, multiplexed protein imaging, could reveal further insights into the spatial regulation of biological processes within tissues. For instance: does the spatial abundance of metabolic enzymes identified by spatial proteomics correlate with the abundance of the metabolites whose reactions they catalyse? Does the spatial location of RNA correlate well with spatial protein abundance, or does this remain poorly correlated? Can high-depth, spatially resolved mass spectrometry-based proteomics be used to inform follow-up imaging mass cytometry, multiplexed ion-beam imaging, and other high-resolution, single-cell multiplexed protein imaging techniques?

To conclude, the results presented within this thesis clearly show the potential for using laser capture microdissection and liquid chromatography-mass spectrometry to investigate the spatial distribution of proteins within tissues in a highly multiplexed manner.

6 Appendix

Supplementary Table 1 – Comparison of optimised method to other spatially resolved proteomic methods.

Reference	Year	Organism	Tissue	Protein IDs	Collection	Protocol	Chromatography	Mass spectrometer	Volume of Tissue	Proteins per 10^3 um3	Proteins per 10^3 um3 per hour
This study	2018	Human	Brain (Molecular Layer)	1521	LCM	SP3	250 nL/min 60 minute gradient	Orbitrap Fusion Lumos	600,000 µm3 (100 * 600 um2 of 10 um thick tissue)	2.535	2.535
This study	2018	Human	Brain (Betz Cells)	3600	LCM	SP3	250 nL/min 60 minute gradient	Orbitrap Fusion Lumos	2,600,000 µm3 (260,000 um2 of 10 um thick tissue)	1.385	1.385
[291]	2013	Rat	Brain	1500	LESA	On-tissue digest	300 nL/min 110 min gradient	Orbitrap Elite	3,530,000 µm3 (670 um diameter on 10 um thick tissue)	0.425	0.232
[292]	2015	Human	Brain	202 to 1773	LCM	Rapigest, In-solution & In Gel	120 minute gradient	Q Exactive	4,000,000 to 80,000,000 µm3 (0.5 mm2 to 10 mm2 of 8 µm thick tissue)	0.0221 to 0.0505	0.0111 to 0.0253
[293]	2015	Human	Brain	303 (Cells) 1144 (Tissue)	LCM	In-solution	30 µl/min 95 min gradient	Q Exactive system	15,000,000 µm3 for cells (1,500,000 µm2 * 10 µm thick tissue), 500,000,000 µm3 for region	0.0202 to 0.02288	0.0127 to 0.0144
[294]	2016	Rat	Brain	1408	LESA	FASP	300 nL/min 100 min gradient	Q-Exactive	15,400,000 µm3 (1280 um diameter area on 12 um thick tissue)	0.091	0.055
[295]	2016	Rat	Brain	250	Laser ablation	FASP	90 min gradient	Amazon ETD Ion Trap	66,000,000 µm3 (1,320,000 um2 of 50 um thick tissue)	0.004	0.003
[136]	2017	Rat	Brain	1027	Hydrogel	hydrogel	300 nL/min 120 min gradient	Orbitrap Fusion	5,690,000 µm3 (777 um diameter hydrogel on 12 um thick tissue)	0.180	0.090
[296]	2017	Rat	Brain	515	LESA	On-tissue digest	250 nL/min 120 min gradient	Q-Exactive	707,000 µm3 (300 um diameter on 10 um thick tissue)	0.728	0.364
[296]	2017	Rat	Brain	700	LESA	On-tissue digest	250 nL/min 120 min gradient	Q-Exactive	707,000 µm3 (300 um diameter on 10 um thick tissue)	0.990	0.495

Reference	Year	Organism	Tissue	Protein IDs	Collection	Protocol	Chromatography	Mass spectrometer	Volume of Tissue	Proteins per 10^3 um3	Proteins per 10^3 um3 per hour
296	2017	Rat	Brain	1303	Parafilm-assisted Microdissection	In-solution	300 nL/min 120 min gradient	Q-Exactive	12,000,000 µm3 (1,000,000 um2 of 12 um thick tissue)	0.109	0.054
297	2017	Mouse	Brain	2170	LCM	SP3	60 minute gradient	Orbitrap Fusion	10,000,000 µm3 (1,000,000 um2 of 10 um thick tissue)	0.217	0.217
298	2017	Human	Brain	2300	LCM	Rapigest	300 nL/min 120 min gradient	Q Exactive Plus	460,000,000 µm3 (11,500,000 um2 * 40 µm thick tissue)	0.005	0.003
212	2018	Rat	Brain	1827	LCM	nanoPOTS	50 nL/min 115 min gradient	Orbitrap Fusion Lumos	370,000 µm3 (200 um diameter of 12 um thick tissue)	4.938	2.576
299	2018	Human	Brain	365	LCM	Rapigest	220 nL/min 85 min gradient	Orbitrap Velos Pro	10,000,000 µm3 (1,000,000 um2 * 10 µm thick tissue)	0.037	0.026
139	2020	Mouse	Liver	2077	LCM	µPOTS	170 min gradient	Orbitrap Fusion Lumos	400,000 µm3 (40,000 um2 of 10 µm thick tissue)	5.193	1.833

Supplementary Table 2 – Two way ANOVA table for SP3/S-Trap digest and 0/1 freeze-thaw cycles on protein groups identified.

Source	DF	Adj SS	Adj MS	F-Value	P-Value
Freeze Thaw Cycles	1	1474	1474	0.05	0.835
Digest	1	1312747	1312747	41.14	0.000
Error	9	287156	31906		
Lack-of-Fit	1	21084	21084	0.63	0.449
Pure Error	8	266072	33259		
Total	11	1601377			

Supplementary Table 3 – Top 20 proteins significant by feature-dependent method only.

		Feature-independent			Feature-dependent		
Gene	**Protein**	**Moran's *I***	***p*-value**	***q*-value**	**Log_2 Fold change**	***p*-value**	***q*-value**
COL11A1	Collagen alpha-1(XI) chain	0.169108	0.003	0.026	3.469157	5.5×10^{-13}	0
LAMC1	Laminin subunit gamma-1	0.188917	0.004	0.031	3.10098	5.3×10^{-11}	0
TUBB3	Tubulin beta-3 chain	0.099693	0.061	0.18	2.609772	3.5×10^{-23}	0
LAMA2	Laminin subunit alpha-2	0.184	0.004	0.031	2.495934	3.5×10^{-7}	3.7×10^{-5}
LAMB2	Laminin subunit beta-2	0.146397	0.011	0.063	2.393519	3.4×10^{-9}	0
DST	Dystonin	0.078762	0.081	0.21	2.319443	2.5×10^{-12}	0
NCAM2	Neural cell adhesion molecule 2	0.026222	0.296	0.43	2.164979	7.0×10^{-09}	0
RCN3	Reticulocalbin-3	0.243977	0.003	0.026	2.144735	6.7×10^{-15}	0
EHD4	EH domain-containing protein 4	0.030633	0.21	0.36	2.043394	1.8×10^{-13}	0
ABCC4	Multidrug resistance-associated protein 4	0.223591	0.002	0.019	1.951451	4.5×10^{-8}	0
AHNAK	Neuroblast differentiation-associated protein AHNAK	0.073524	0.087	0.21	-1.00473	6.5×10^{-4}	0.020
PDIA5	Protein disulfide-isomerase A5	0.21941	0.002	0.019	-1.00662	3.5×10^{-09}	0
SERPINB8	Serpin B8	0.348612	0.006	0.042	-1.03377	5.5×10^{-4}	0.018
GDAP1	Ganglioside-induced differentiation-associated protein 1	0.472938	0.015	0.076	-1.14956	2.5×10^{-4}	0.0098
ANXA2	Annexin A2	0.152119	0.005	0.037	-1.29518	5.1×10^{-14}	0
SLC2A1	Solute carrier family 2, facilitated glucose transporter member 1	0.00653	0.352	0.48	-1.53484	9.3×10^{-4}	0.026
SYNE1	Nesprin-1	0.215476	0.006	0.042	-1.88157	4.6×10^{-6}	3.3×10^{-4}
TPD52L1	Tumor protein D53	0.359185	0.004	0.031	-2.21012	2.0×10^{-7}	2.0×10^{-5}
SPTB	Spectrin beta chain, erythrocytic	0.140109	0.023	0.098	-2.67892	2.5×10^{-3}	0.049
SPTA1	Spectrin alpha chain, erythrocytic 1	0.125613	0.027	0.11	-2.97088	9.9×10^{-4}	0.027

Supplementary Table 4 – Top 20 proteins significant by feature-independent method only.

		Feature-independent			Feature-dependent		
Gene	**Protein**	**Moran's *I***	***p*-value**	***q*-value**	**Log_2 Fold change**	***p*-value**	***q*-value**
HGS	Hepatocyte growth factor-regulated tyrosine kinase substrate	0.696407	0.001	0.011106	0.628177	0.0077	0.097
ZRANB2	Zinc finger Ran-binding domain-containing protein 2	0.59069	0.001	0.011106	-0.63119	9.8×10^{-4}	0.026
AGPAT5	1-acyl-sn-glycerol-3-phosphate acyltransferase epsilon	0.556017	0.001	0.011106	-0.97007	2.8×10^{-7}	1.9×10^{-5}
OPTN	Optineurin	0.552742	0.001	0.011106	-0.1167	0.64	0.88
PHPT1	14 kDa phosphohistidine phosphatase	0.500926	0.001	0.011106	0.673814	0.0016	0.036
CRP	C-reactive protein	0.486552	0.001	0.011106	0.663102	0.13	0.41
SHPK	Sedoheptulokinase	0.469403	0.001	0.011106	-0.20617	0.18	0.48
MARK1	Serine/threonine-protein kinase MARK1	0.468411	0.001	0.011106	-0.70774	5.4×10^{-6}	4.0×10^{-7}
HINT3	Histidine triad nucleotide-binding protein 3	0.461732	0.001	0.011106	-0.23322	0.18	0.47
SDF4	45 kDa calcium-binding protein	0.453182	0.001	0.011106	0.763209	1.3×10^{-6}	7.9×10^{-5}
GGCX	Vitamin K-dependent gamma-carboxylase	0.452601	0.001	0.011106	-0.10696	0.55	0.83
NUTF2	Nuclear transport factor 2	0.452005	0.001	0.011106	0.891433	1.0×10^{-6}	6.6×10^{-6}
CALM1	Calmodulin	0.449213	0.001	0.011106	0.821243	3.1×10^{-6}	1.8×10^{-4}
HEBP1	Heme-binding protein 1	0.447967	0.001	0.011106	-0.84338	1.4×10^{-9}	0
MOB1A;MOB1B	MOB kinase activator 1A/1B	0.446075	0.001	0.011106	-0.01163	0.95	1
RUFY3	Protein RUFY3	0.442192	0.001	0.011106	-0.73435	3.2×10^{-7}	1.9×10^{-5}
LIG3	DNA ligase 3	0.441895	0.001	0.011106	-0.76606	4.4×10^{-8}	0
DAG1	Dystroglycan	0.441259	0.001	0.011106	-0.57653	3.5×10^{-5}	0.0023
WDR91	WD repeat-containing protein 91	0.437029	0.001	0.011106	0.57018	0.094	0.35
SIRT3	NAD-dependent protein deacetylase sirtuin-3, mitochondrial	0.436094	0.001	0.011106	-0.18816	0.23	0.54

Supplementary Table 5 – Top 20 proteins significant by feature-dependent and feature-independent methods.

		Feature-independent			Feature-dependent		
Gene	**Protein**	**Moran's *I***	***p*-value**	***q*-value**	**Log_2 Fold change**	***p*-value**	***q*-value**
GFAP	Glial fibrillary acidic protein	0.621867	0.001	0.011106	-2.33552	5.9×10^{-8}	0
NEGR1	Neuronal growth regulator 1	0.606709	0.001	0.011106	-2.57021	3.1×10^{-18}	0
PRPH	Peripherin	0.581984	0.001	0.011106	-7.03626	1.4×10^{-18}	0
FMNL2	Formin-like protein 2	0.578553	0.001	0.011106	-2.85385	2.0×10^{-13}	0
SORT1	Sortilin	0.527718	0.001	0.011106	-2.76916	3.4×10^{-11}	0
PRCP	Lysosomal Pro-X carboxypeptidase	0.472838	0.001	0.011106	-3.11934	6.1×10^{-17}	0
FAM129A	Protein Niban	0.454052	0.001	0.011106	-3.08433	9.1×10^{-9}	0
GRID2	Glutamate receptor ionotropic, delta-2	0.453626	0.001	0.011106	-2.37325	8.1×10^{-12}	0
CKB	Creatine kinase B-type	0.435337	0.001	0.011106	-2.79793	1.90×10^{-14}	0
CLIC4	Chloride intracellular channel protein 4	0.421645	0.001	0.011106	3.105953	1.8×10^{-16}	0
PYGL	Glycogen phosphorylase, liver form	0.404599	0.001	0.011106	4.470298	4.4×10^{-21}	0
FABP7	Fatty acid-binding protein, brain	0.387561	0.001	0.011106	-2.52139	2.1×10^{-24}	0
SRGAP3	SLIT-ROBO Rho GTPase-activating protein 3	0.387351	0.001	0.011106	2.956308	1.6×10^{-17}	0
ANXA1	Annexin A1	0.351926	0.001	0.011106	2.708949	4.8×10^{-15}	0
CRABP1	Cellular retinoic acid-binding protein 1	0.34203	0.001	0.011106	3.012004	2.8×10^{-15}	0
CSPG4	Chondroitin sulfate proteoglycan 4	0.332184	0.001	0.011106	2.879803	1.1×10^{-10}	0
COL2A1	Collagen alpha-1(II) chain	0.307641	0.001	0.011106	5.209607	2.2×10^{-15}	0
NEBL	Nebulette	0.278312	0.001	0.011106	2.685355	1.0×10^{-13}	0
NID1	Nidogen-1	0.269942	0.001	0.011106	2.88664	1.3×10^{-05}	0.001027
ASPH	Aspartyl/asparaginyl beta-hydroxylase	0.219415	0.001	0.011106	3.514204	2.1×10^{-25}	0

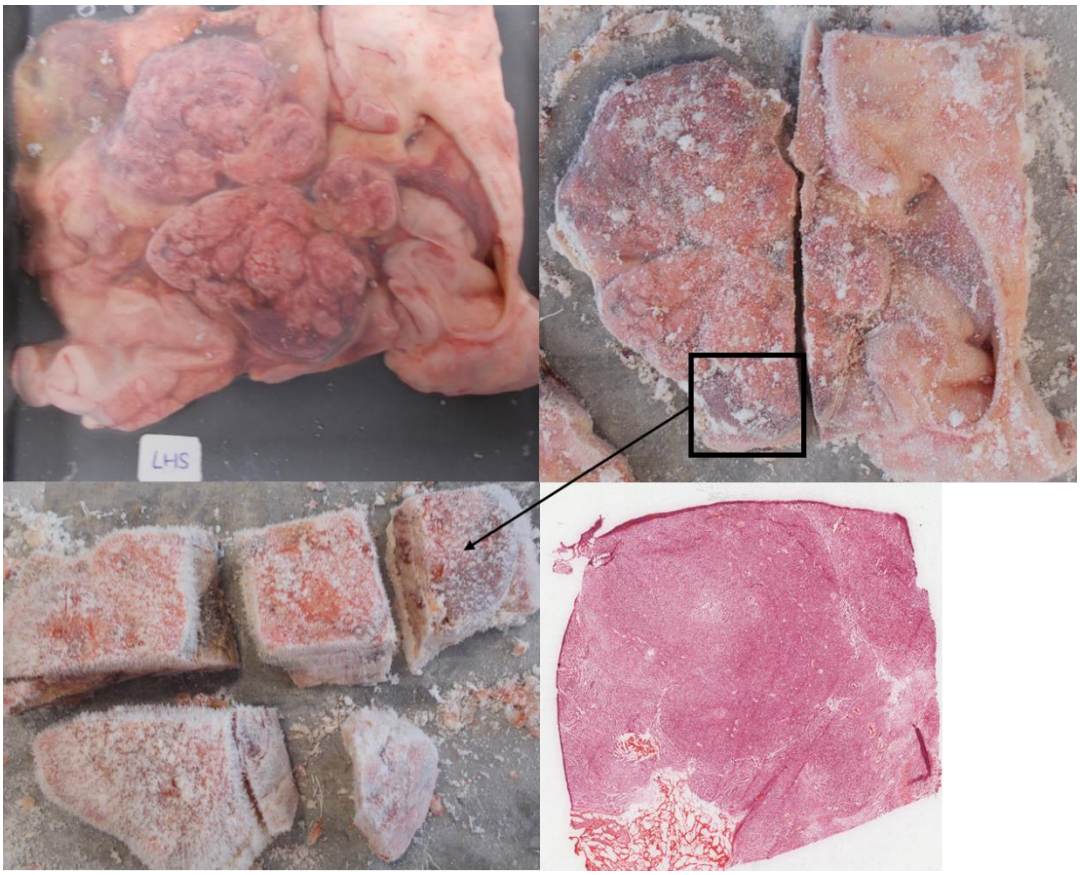

Figure S1 – Gross dissection of ATRT tumour

The P2 coronal slice (upper left) was taken and the tumour was mostly separated from surrounding tissue (upper right). The tumour was split into five pieces (lower left) and each was H&E stained and inspected. The upper right block (lower right) was selected for further use based on quality of preserved morphology.

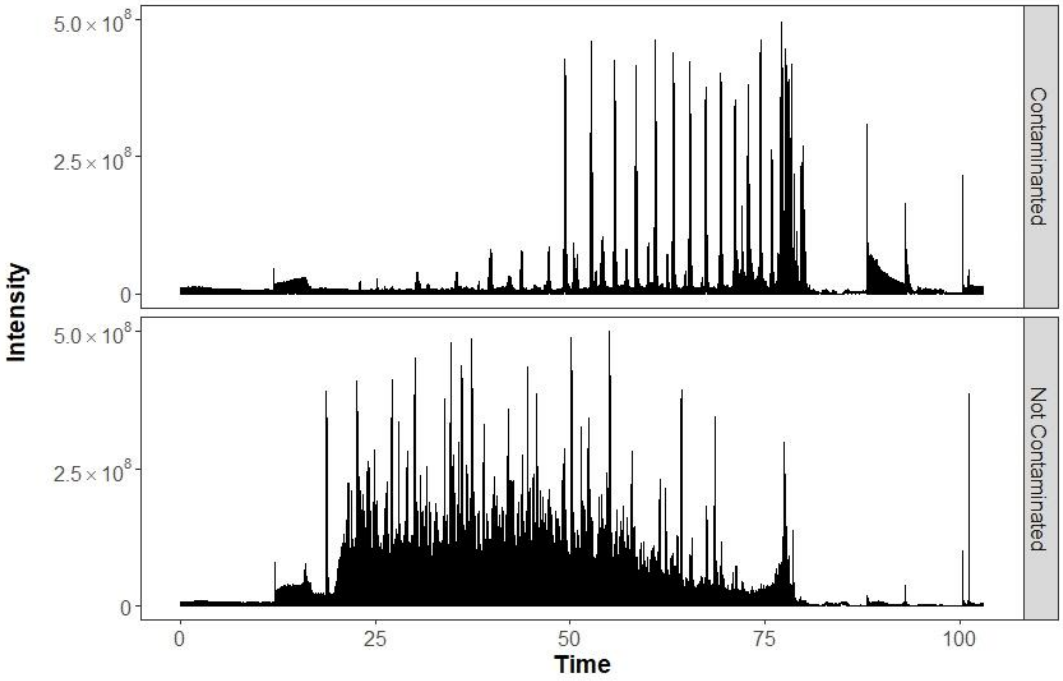

Figure S2 – Example total ion chromatogram of PEG contaminated sample

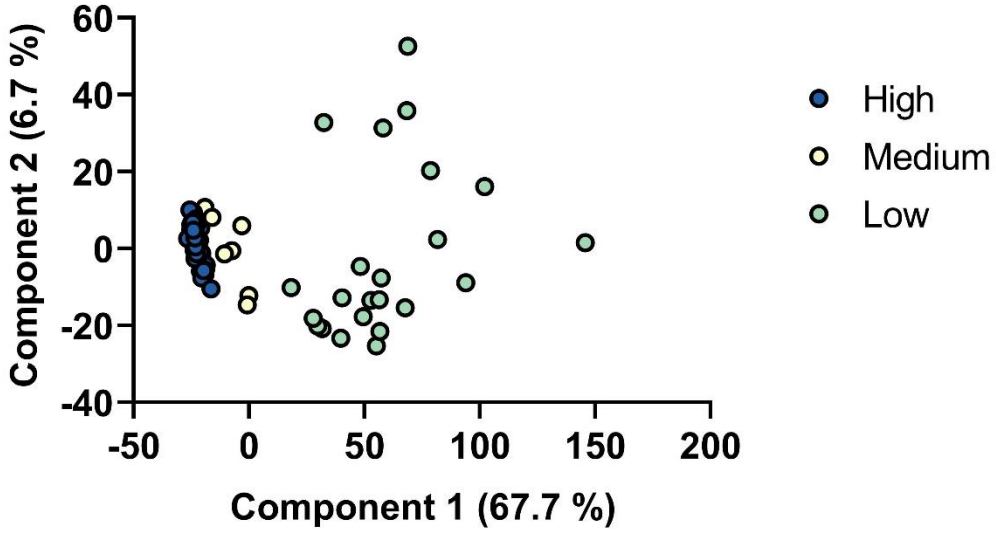

Figure S3 – PCA plot of spatial proteomics data from ATRT Slide 1

Principal component analysis shows separation of proteomic data based on quality of sample. High, Medium and Low refer to amount of quantified proteins as in Figure 26.

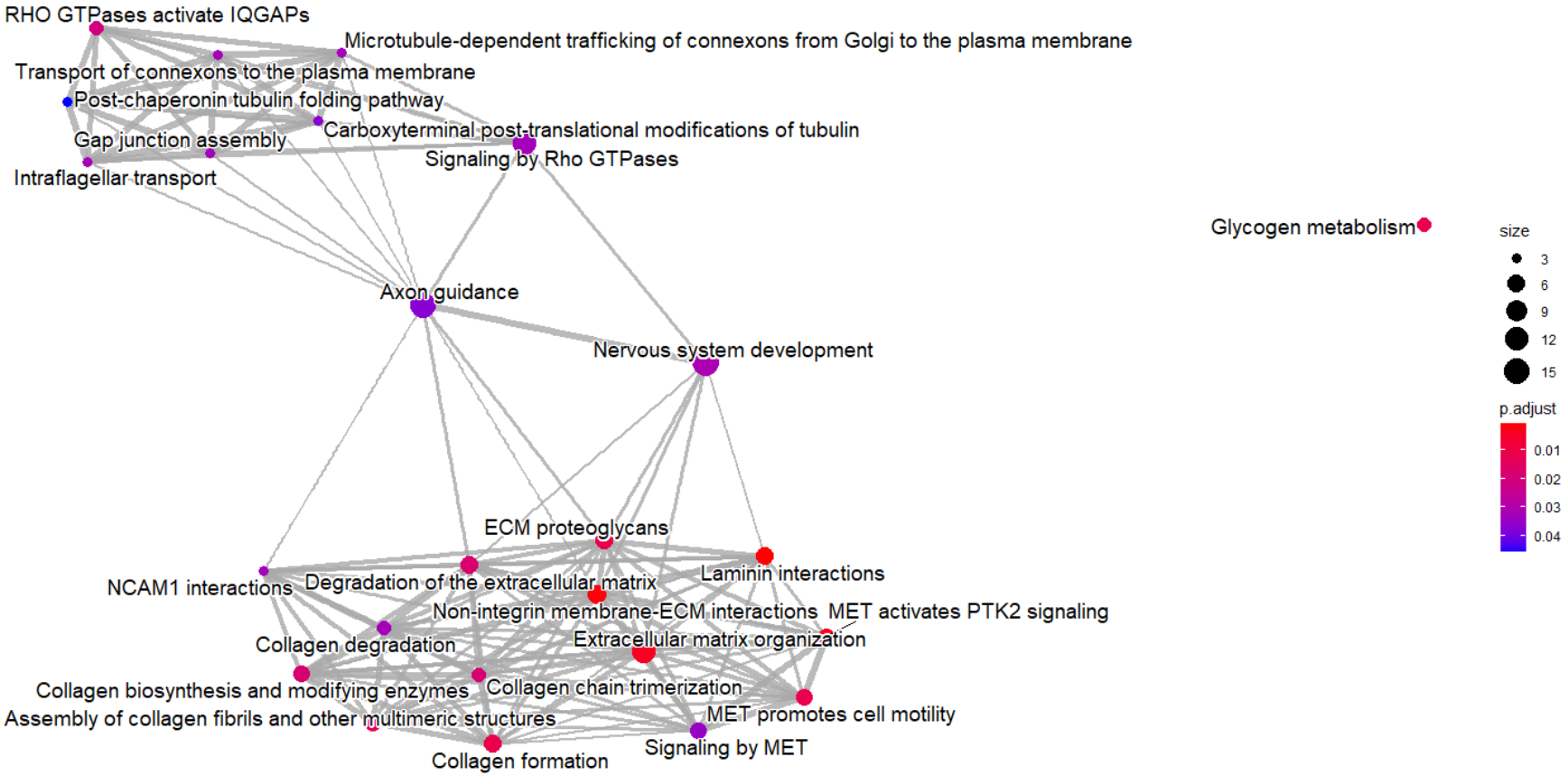

Figure S4 – Reactome pathways up-regulated in solid tumour

Functional enrichment of Reactome pathway terms within differentially expressed proteins between the region of solid tumour and brain/tumour interface. Size refers to the number of differential proteins annotated with that term; adj. p-value refers to Benjamini-Hochberg corrected p-values. Edges represent shared protein members between annotations.

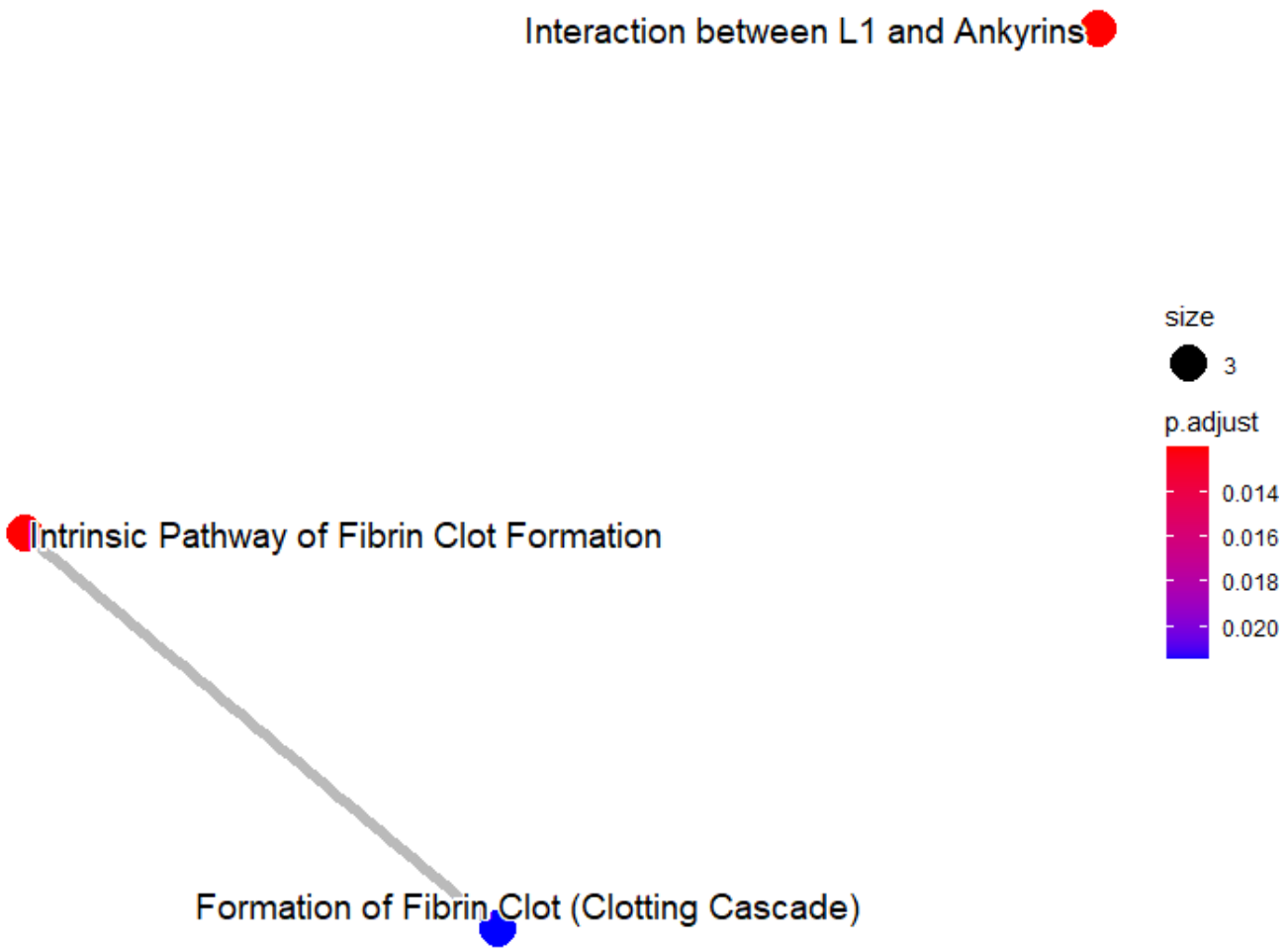

Figure S5 - Reactome pathways down-regulated in solid tumour

Functional enrichment of Reactome pathway terms within differentially expressed proteins between the region of solid tumour and brain/tumour interface. Size refers to the number of differential proteins annotated with that term; adj. p-value refers to Benjamini-Hochberg corrected p-values. Edges represent shared protein members between annotations.

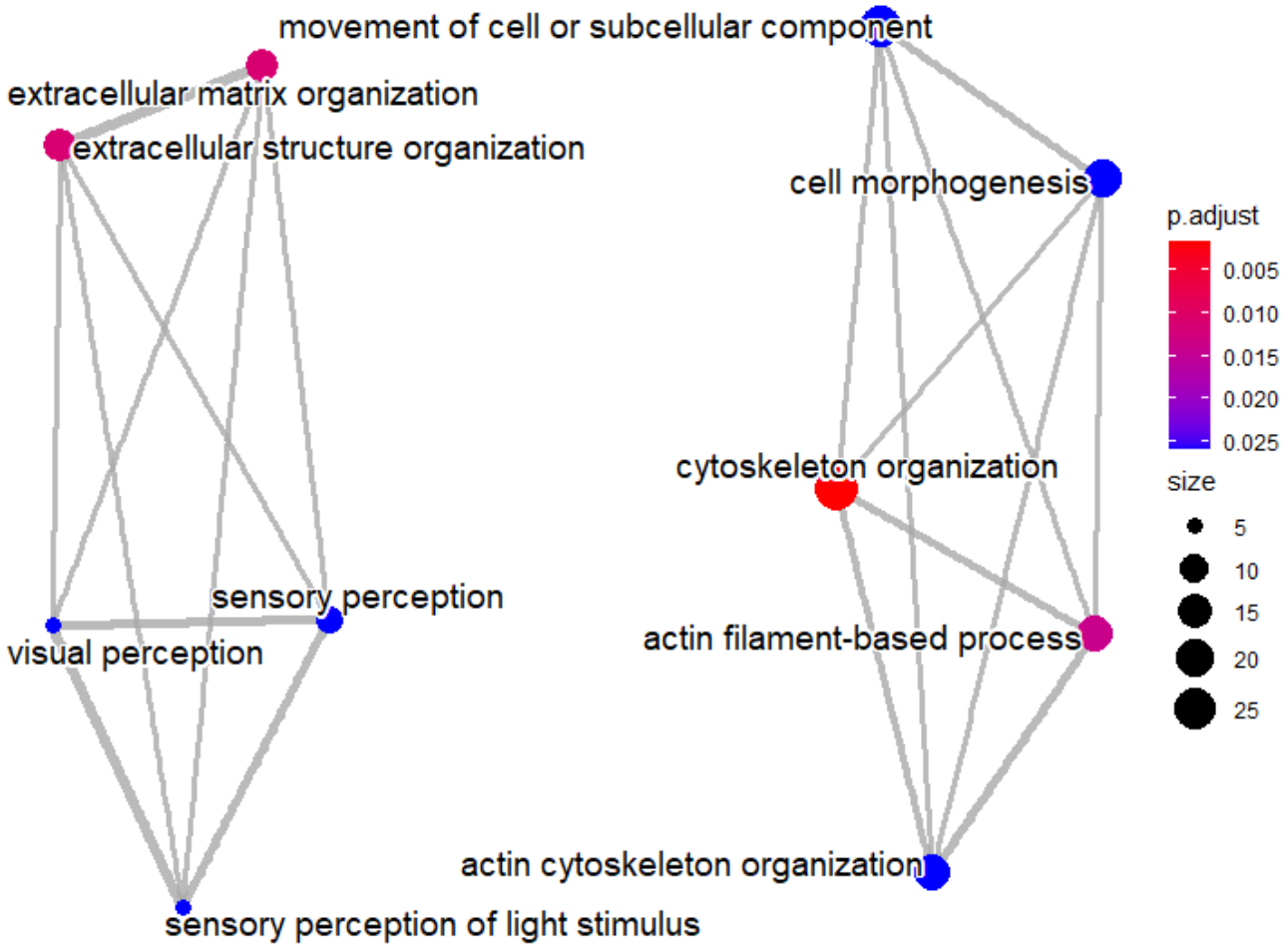

Figure S6 – Gene Ontology biological process terms up-regulated in solid tumour

Functional enrichment of Gene Ontology biological process terms within differentially expressed proteins between the region of solid tumour and brain/tumour interface. Size refers to the number of differential proteins annotated with that term; adj. p-value refers to Benjamini-Hochberg corrected p-values. Edges represent shared protein members between annotations.

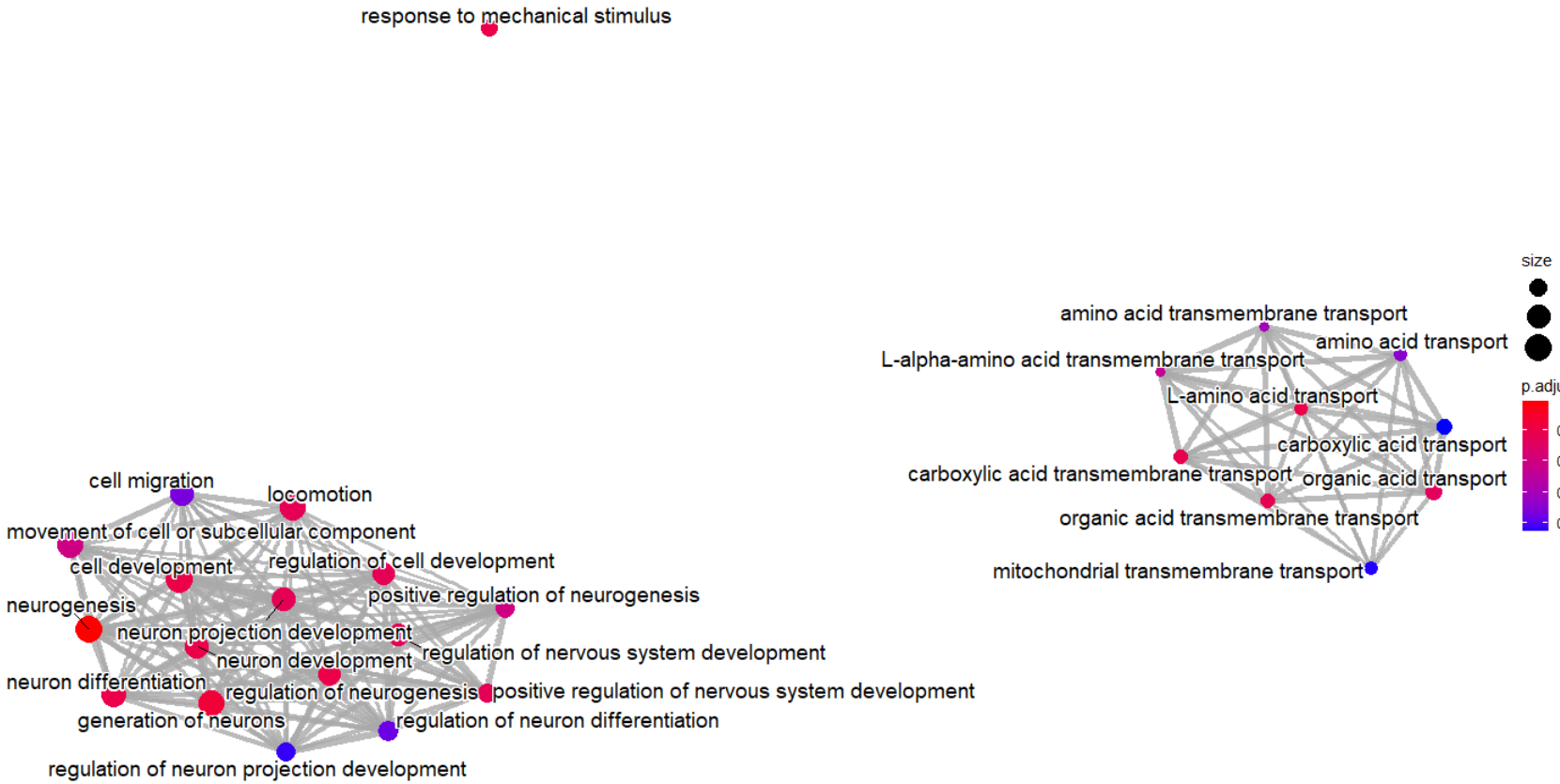

Figure S7 – Gene Ontology biological process terms down-regulated in solid tumour

Functional enrichment of Gene Ontology biological process terms within differentially expressed proteins between the region of solid tumour and brain/tumour interface. Size refers to the number of differential proteins annotated with that term; adj. p-value refers to Benjamini-Hochberg corrected p-values. Edges represent shared protein members between annotations.

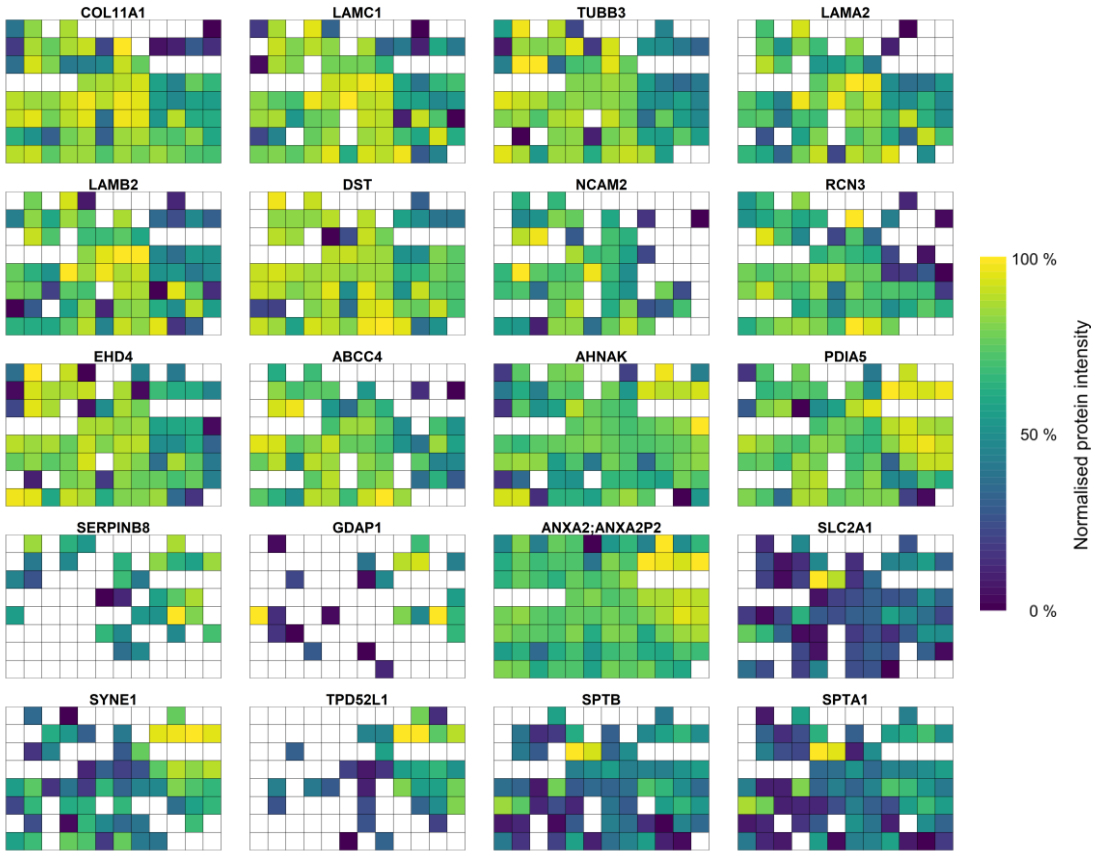

Figure S8 – Top 20 significant proteins by the feature-dependent method only

Normalised protein intensity of the 20 proteins with the largest $\log_2$ fold-changes that were called as significant by the feature-dependent method only. Corresponds to Supplementary Table 3. Normalised protein intensities are scaled separately for each protein. White = not quantified.

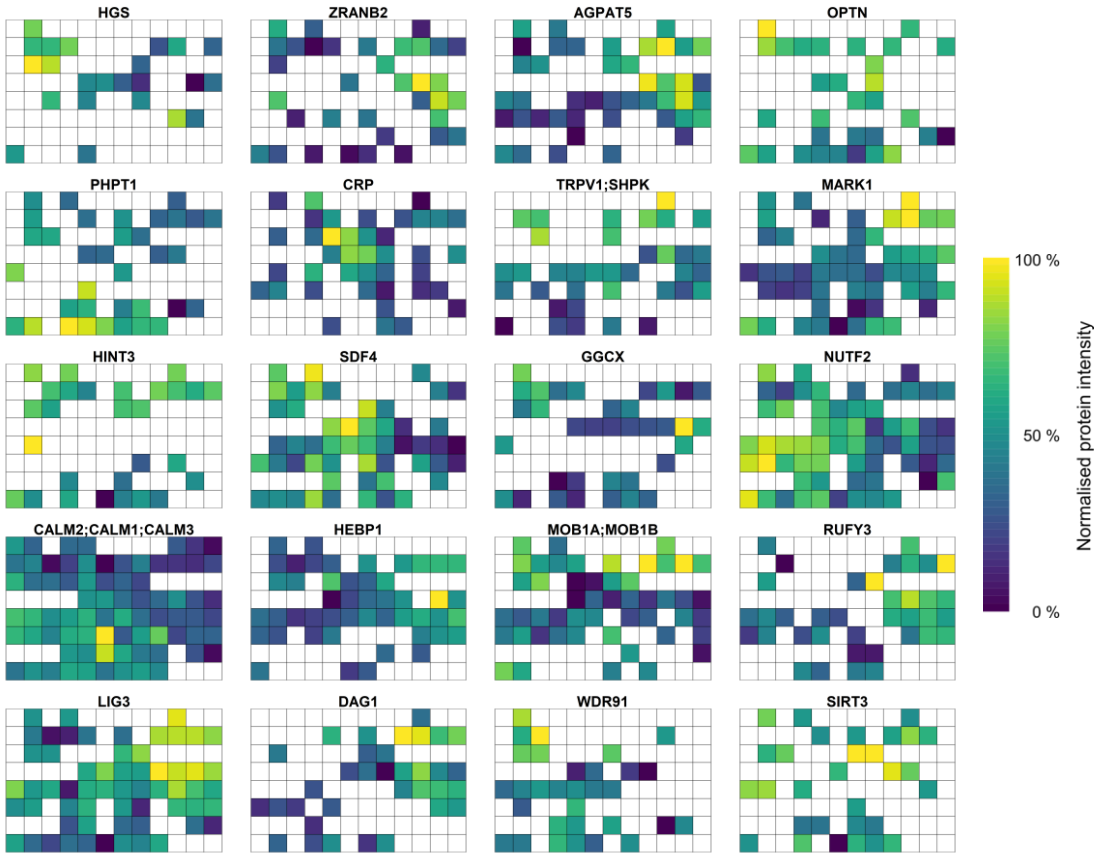

Figure S9 – Top 20 significant proteins by the feature-independent method only

Normalised protein intensity of the 20 proteins with the largest Moran's *I* that were called as significant by the feature-independent method only. Corresponds to Supplementary Table 4. Normalised protein intensities are scaled separately for each protein. White = not quantified.

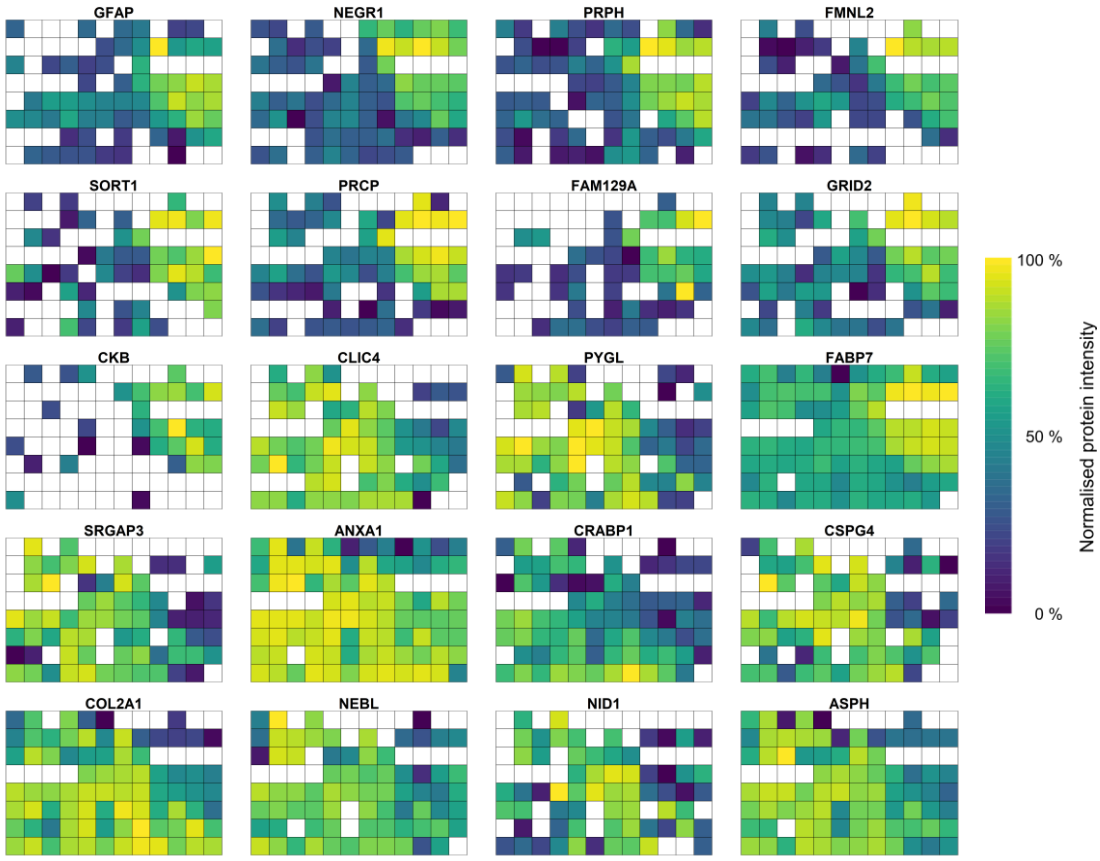

Figure S10 – Top 20 significant proteins by the feature-dependent and feature-independent methods

Normalised protein intensity of the top 20 proteins with the largest $\log_2$ fold-changes that were called as significant by the both feature-dependent and feature-independent methods. Corresponds to Supplementary Table 5. Normalised protein intensities are scaled separately for each protein. White = not quantified.

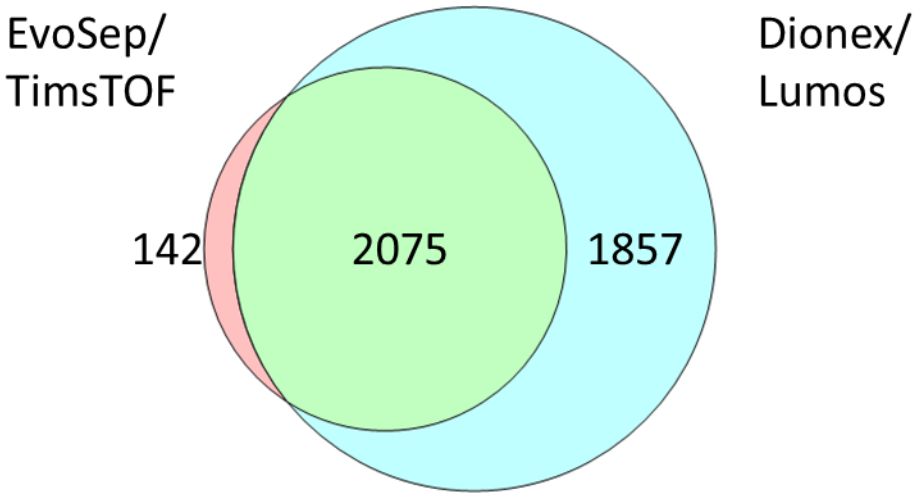

Figure S11 – Overlap in protein identifications between the two LC-MS/MS systems

Proteins identified in at least one sample of the first raster experiment. EvoSep/TimsTOF uses 11.5 min linear gradients, Dionex/Lumos uses 60 minute linear gradients.

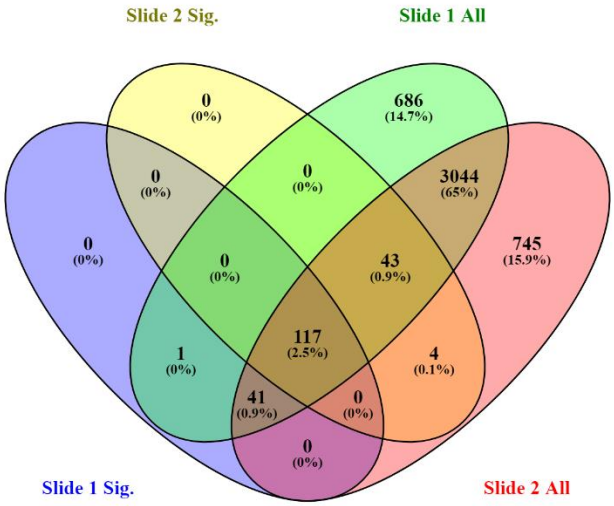

Figure S12 – Overlap of differentially expressed proteins determined by the feature-guided method

Slide 1 All – All proteins detected within the first raster experiment. Slide 1 Sig. – Proteins from the first raster experiment called as significant by a two-sample *t*-test between pixels of solid tumour and brain/tumour interface that also have a $\log_2$ fold-change of at least ± 1. Slide 2 All – All proteins detected within the second raster experiment. Slide 2 Sig. – Proteins from the second raster experiment called as significant by a two-sample *t*-test between pixels of solid tumour and brain/tumour interface that also have a $\log_2$ fold-change of at least ± 1.

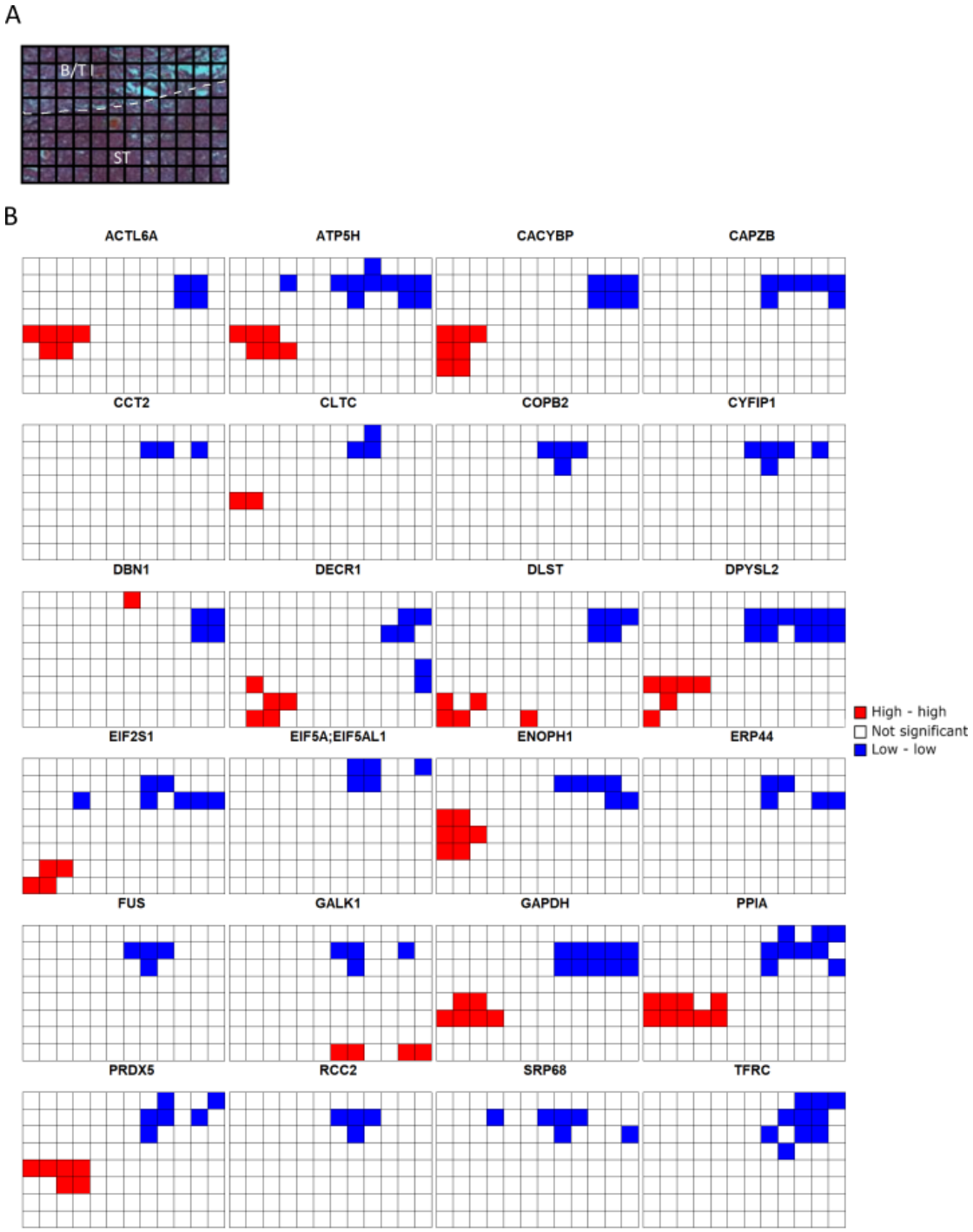

Figure S13 – Examples of low-low regions that correlate with area of low tissue density

(A) Composite image of H&E stained ATRT section stitched from multiple fields of view acquired directly on the LCM instrument, with raster sampling LCM annotations overlaid. Dashed line represents transfer of the pathologist's annotation to this section. BTI: brain/tumour interface. BV: blood vessel. ST: solid tumour. (B) Pixel classification by local Moran's *I* analysis of 24 example proteins.

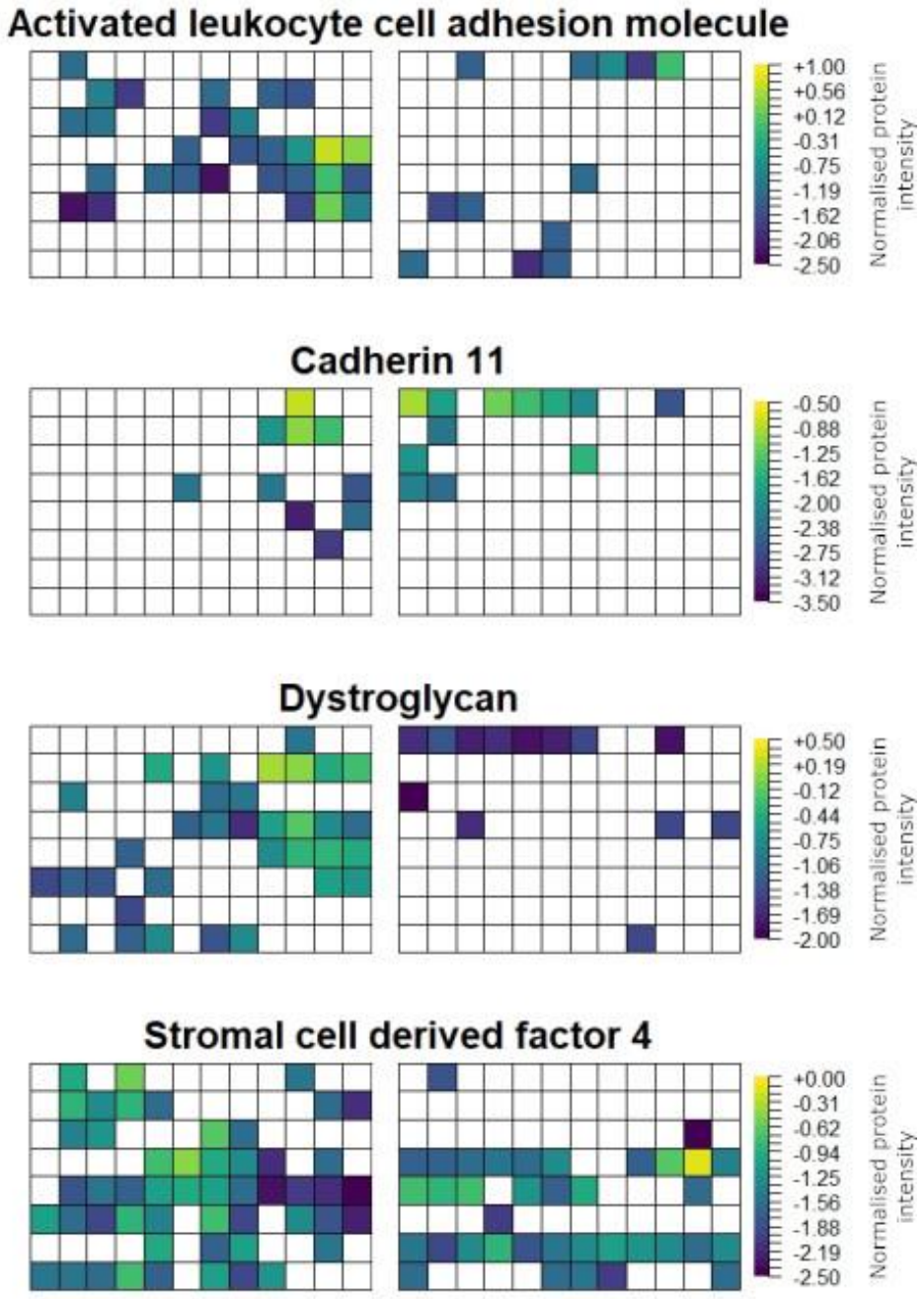

Figure S14 – Spatial proteomic maps of example proteins significant by feature-independent analysis in experiment one only

Normalised protein intensity maps of four example proteins called as significant by the feature-independent Moran's *I* analysis only in the first experiment (left maps) and not the second experiment (right maps). White = not quantified.

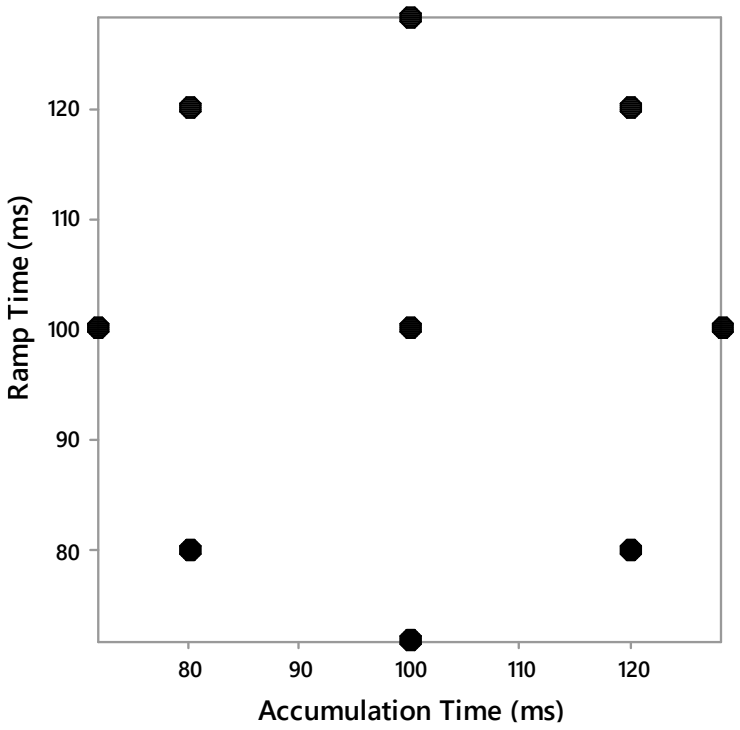

Figure S15 – Parameter space tested in TIMS cell optimisation

Black circles mark parameter combinations tested to determine optimal settings.

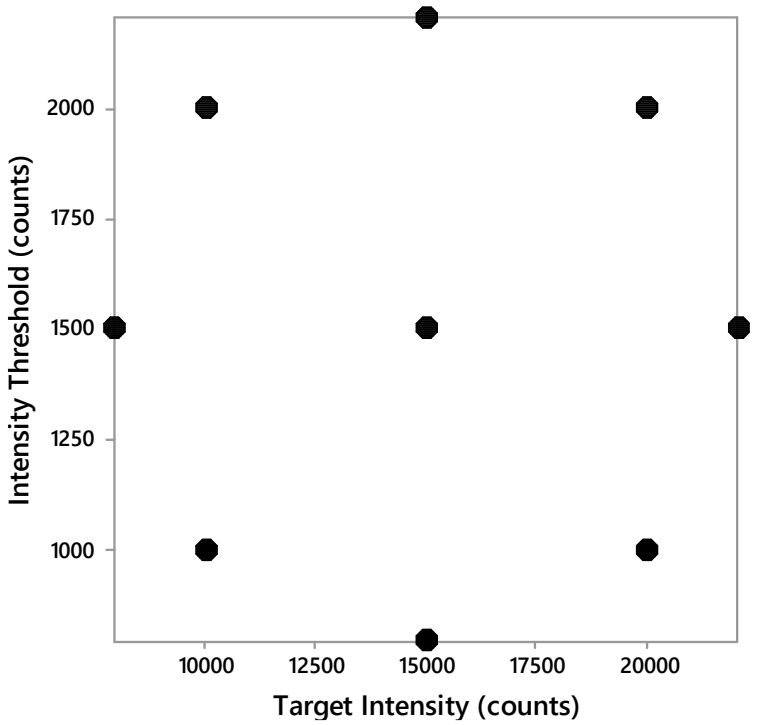

Figure S16 – Parameter space tested in fragmentation intensity optimisation

Black circles mark parameter combinations tested to determine optimal settings.

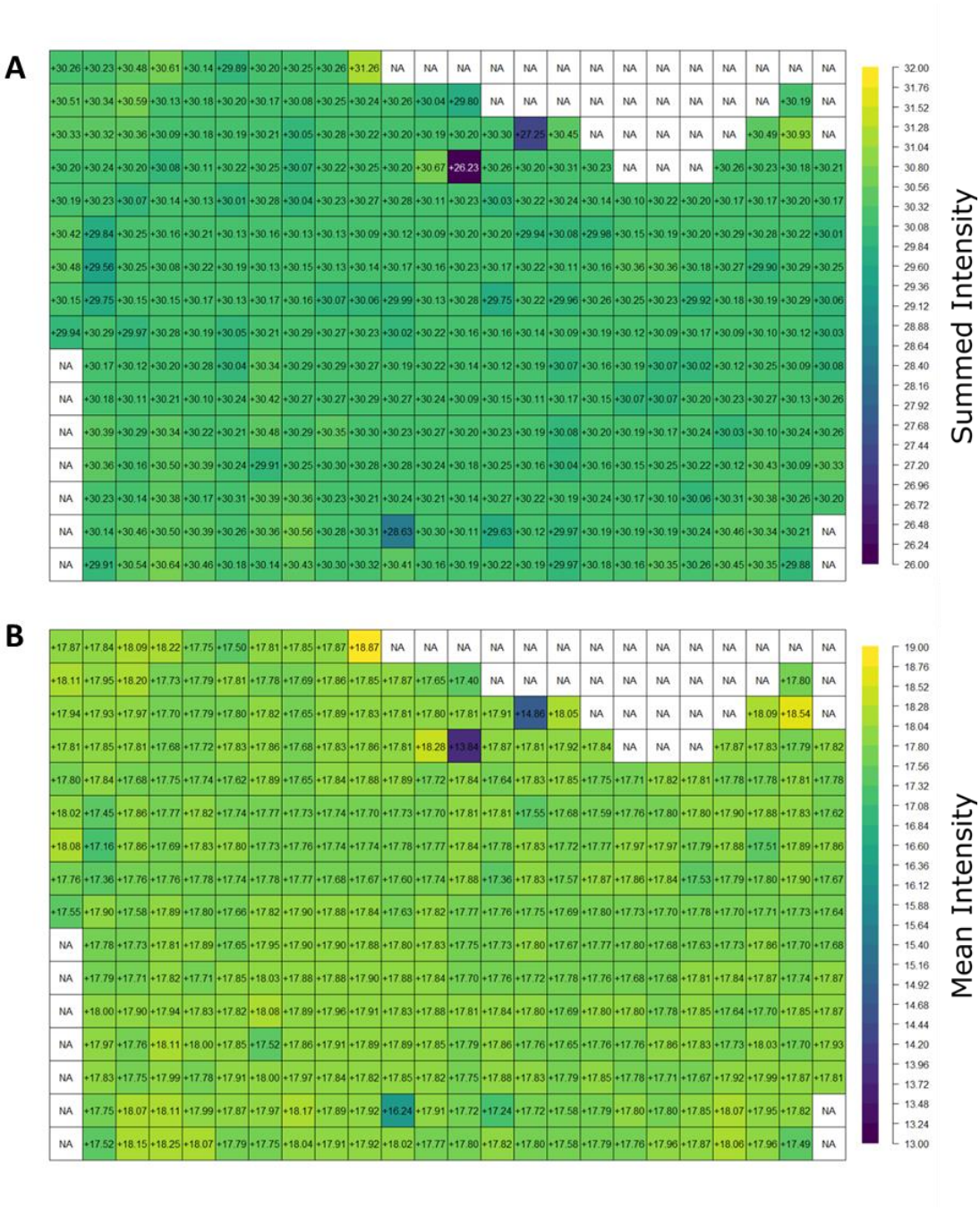

Figure S17 - Summed and mean intensities of each sample with pixel values overlaid

Maps of the $\log_2$ transformed (A) summed and (B) mean intensities of each sample. Samples corresponding to haemorrhage and empty pixels were removed (white).

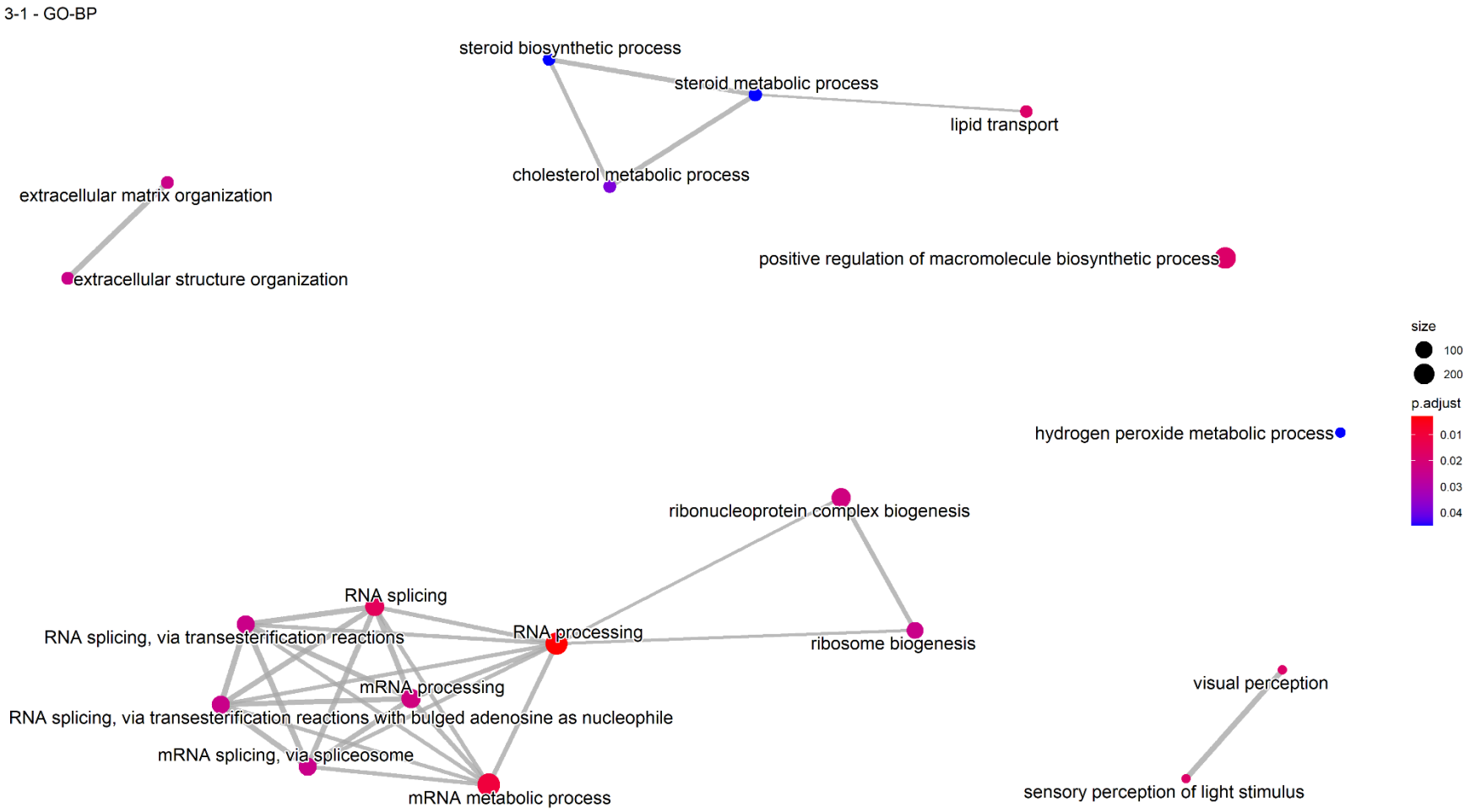

Figure S18 – Functional analysis of Clusters 1 and 3

Gene set enrichment analysis of Gene Ontology Biological process terms with clusters 1 & 2 as defined by hierarchical clustering and dynamic tree cutting. Size refers to the number of differential proteins annotated with that term; p.adjust refers to Benjamini-Hochberg corrected p-values. Edges represent shared protein members between annotations.

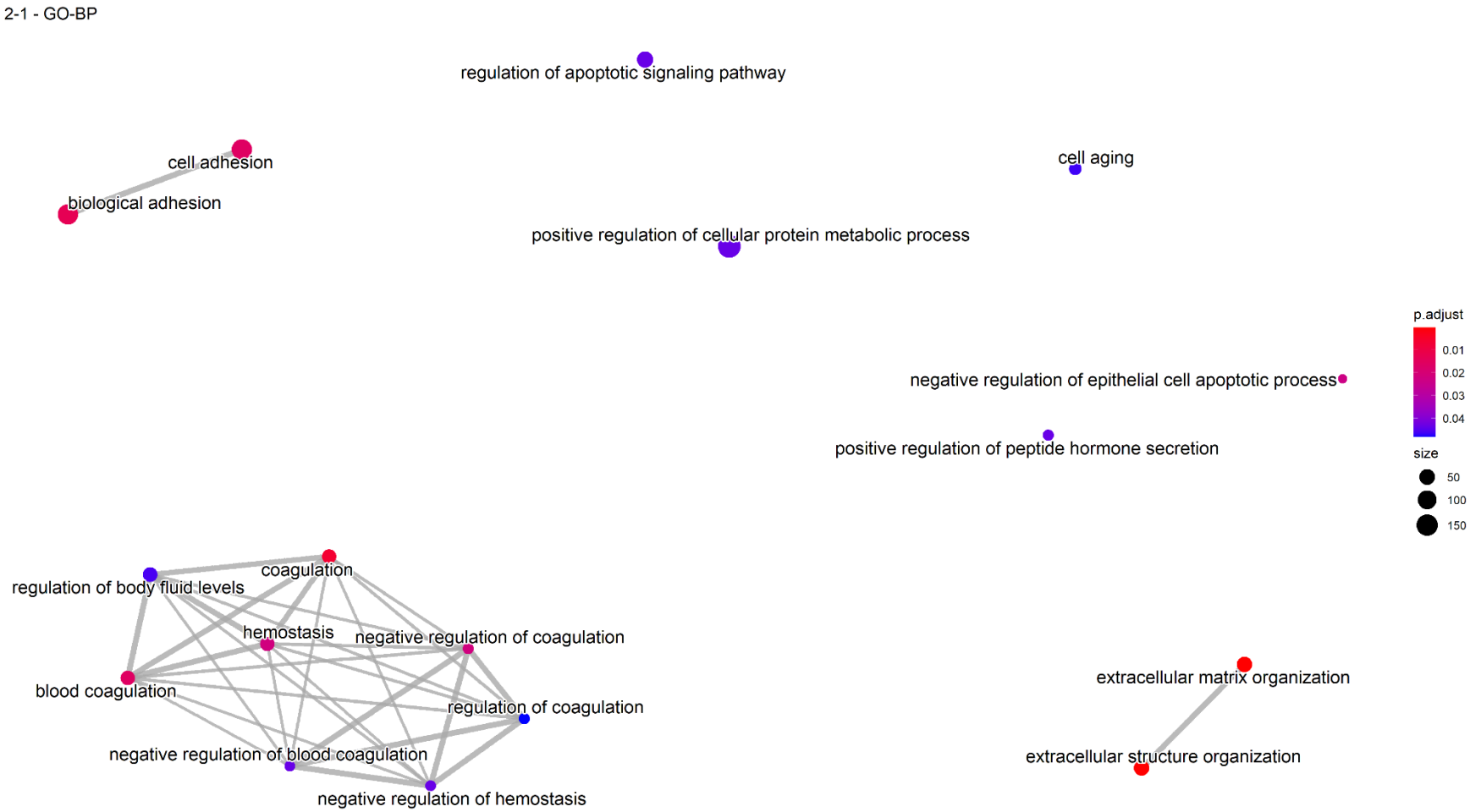

Figure S19 – Functional analysis of Clusters 1 and 2

Gene set enrichment analysis of Gene Ontology Biological process terms with clusters 1 & 3 as defined by hierarchical clustering and dynamic tree cutting. Size refers to the number of differential proteins annotated with that term; p.adjust refers to Benjamini-Hochberg corrected p-values. Edges represent shared protein members between annotations.

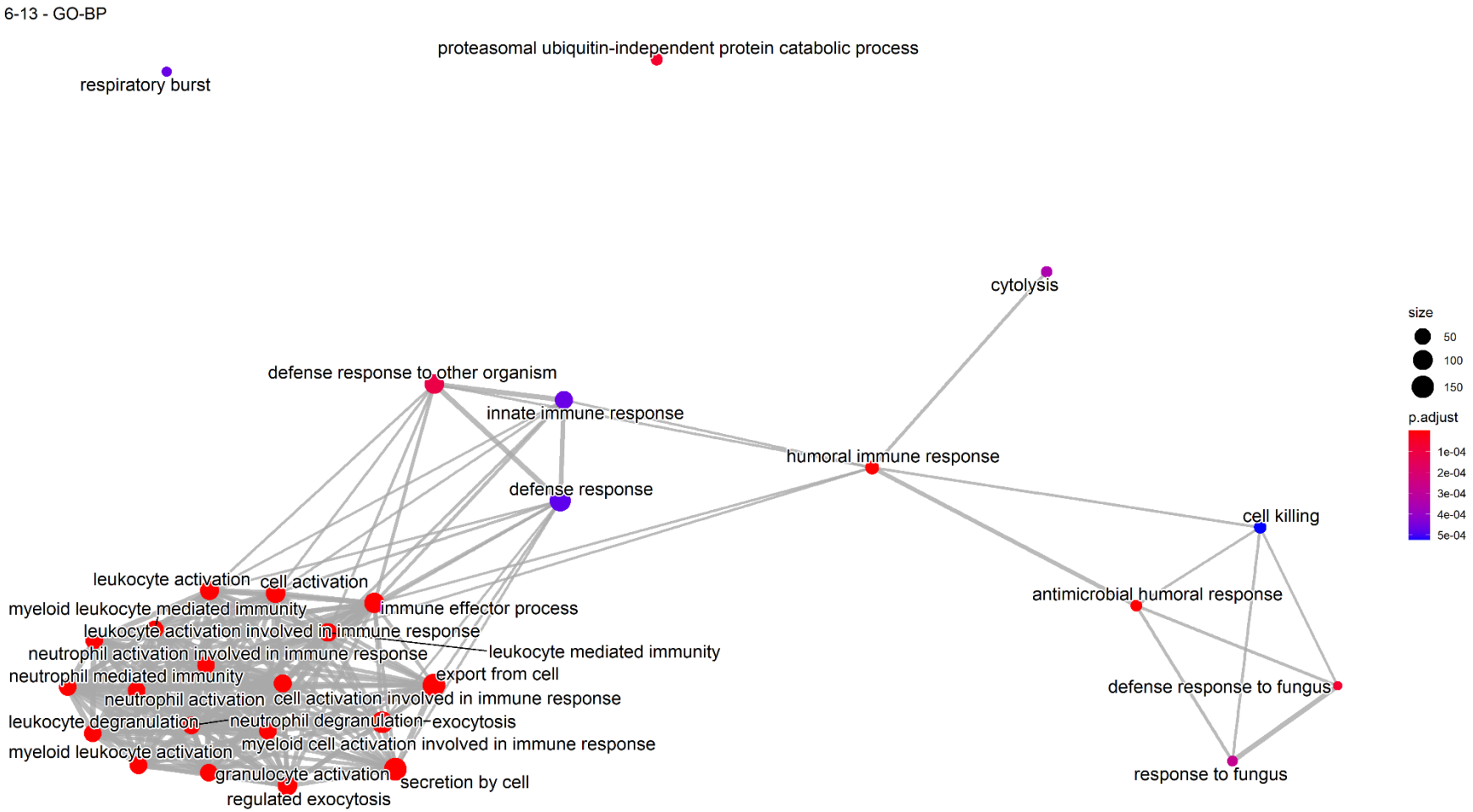

Figure S20 – Functional analysis of Clusters 6 and 13

Gene set enrichment analysis of Gene Ontology Biological process terms with clusters 1 & 3 as defined by hierarchical clustering and dynamic tree cutting. Size refers to the number of differential proteins annotated with that term; p.adjust refers to Benjamini-Hochberg corrected p-values. Edges represent shared protein members between annotations.

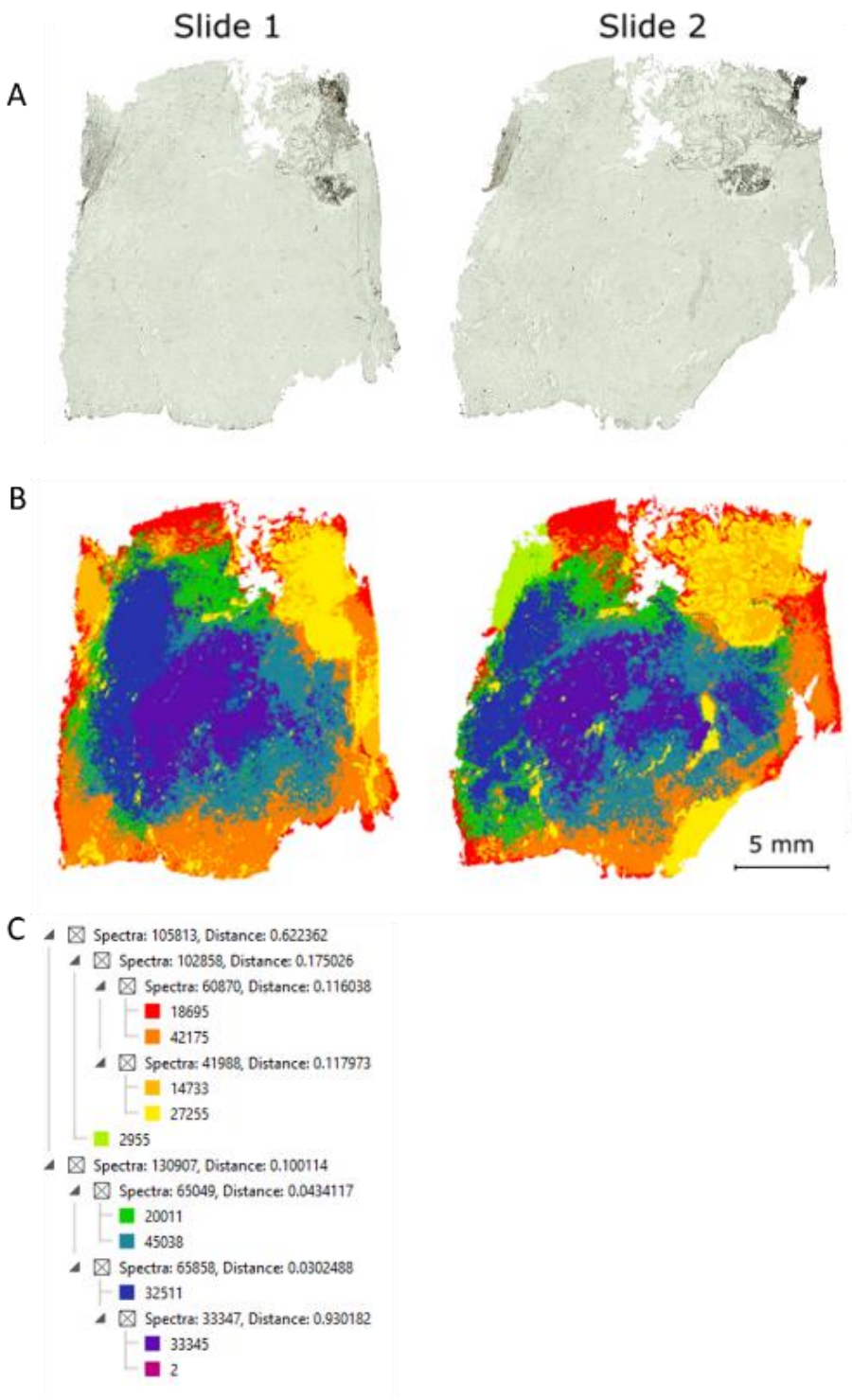

Figure S21 – MALDI imaging of ATRT peptides

Two adjacent ATRT tissue sections were sprayed with trypsin to digest proteins *in-situ* and then imaged by MALDI imaging. (A) Bisecting K-means segmentation of 498 m/z intervals. Segmentation maps were split to resemble the ATRT tissue block histopathology. (C) Segmentation tree clusters, colours within each node are most similar to each other.

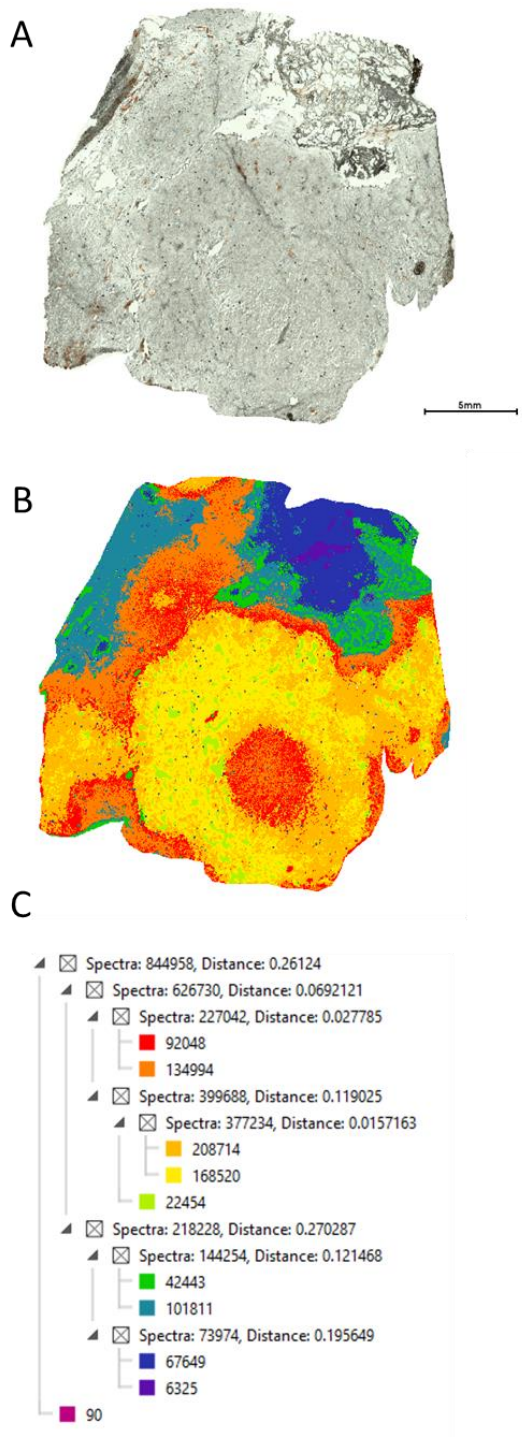

Figure S22 – MALDI imaging of ATRT lipids

(A) Brightfield image of unstained ATRT tissue prior to MALDI imaging. (B) Bisecting K-means segmentation of 498 m/z intervals. Segmentation maps were split to resemble the ATRT tissue block histopathology. (C) Segmentation tree clusters, colours within each node are most similar to each other.

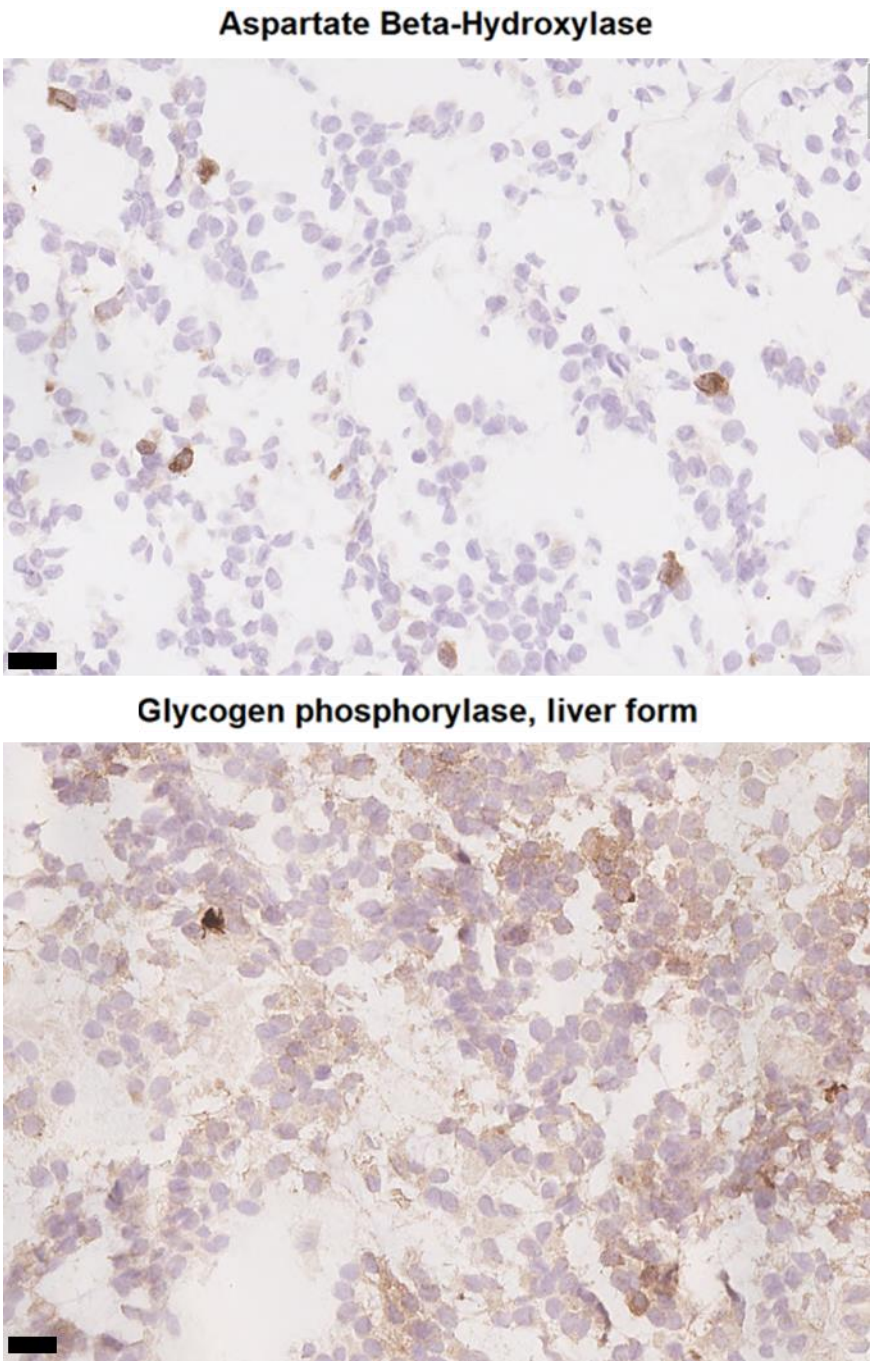

Figure S23 – Higher magnification of brain/tumour interface PYGL and ASPH immunohistochemistry

Representative region of the brain/tumour interface from PYGL and ASPH IHC images. Scale bar = 20 µm.

7 References

1. Altelaar, A. F. M., Munoz, J. & Heck, A. J. R. Next-generation proteomics: Towards an integrative view of proteome dynamics. *Nature Reviews Genetics* vol. 14 35–48 (2013).

2. Karas, M. & Hillenkamp, F. Laser Desorption Ionization of Proteins with Molecular Masses Exceeding 10 000 Daltons. *Analytical Chemistry* vol. 60 2299–2301 (1988).

3. Fenn, J. B., Mann, M., Meng, C. K., Wong, S. F. & Whitehouse, C. M. Electrospray ionization for mass spectrometry of large biomolecules. *Science* vol. 246 64–71 (1989).

4. Davis, S. *et al.* Expanding Proteome Coverage with CHarge Ordered Parallel Ion aNalysis (CHOPIN) Combined with Broad Specificity Proteolysis. *J. Proteome Res.* **16**, 1288–1299 (2017).

5. Hebert, A. S. *et al.* The one hour yeast proteome. *Mol. Cell. Proteomics* **13**, 339–47 (2014).

6. Bekker-Jensen, D. B. *et al.* An Optimized Shotgun Strategy for the Rapid Generation of Comprehensive Human Proteomes. *Cell Syst.* **11**, 214–221 (2017).

7. Meier, F., Geyer, P. E., Virreira Winter, S., Cox, J. & Mann, M. BoxCar acquisition method enables single-shot proteomics at a depth of 10,000 proteins in 100 minutes. *Nat. Methods* **15**, 440–448 (2018).

8. Demichev, V. *et al.* High sensitivity dia-PASEF proteomics with DIA-NN and FragPipe. *bioRxiv* 2021.03.08.434385 (2021) doi:10.1101/2021.03.08.434385.

9. Aebersold, R. & Mann, M. Mass-spectrometric exploration of proteome structure and function. *Nature* **537**, 347–355 (2016).

10. Pandey, A. & Mann, M. Proteomics to study genes and genomes. *Nature* vol. 405 837–846 (2000).

11. Weston, L. A., Bauer, K. M. & Hummon, A. B. Comparison of bottom-up proteomic approaches for LC-MS analysis of complex proteomes. *Anal. Methods* **5**, 4615–4621 (2013).

12. Burkhart, J. M., Schumbrutzki, C., Wortelkamp, S., Sickmann, A. & Zahedi, R. P. Systematic and quantitative comparison of digest efficiency and specificity reveals the impact of trypsin quality on MS-based proteomics. *J. Proteomics* **75**, 1454–1462 (2012).

13. Olsen, J. V., Ong, S. E. & Mann, M. Trypsin cleaves exclusively C-terminal to arginine and lysine residues. *Mol. Cell. Proteomics* **3**, 608–614 (2004).

14. Guo, X., Trudgian, D. C., Lemoff, A., Yadavalli, S. & Mirzaei, H. Confetti: A Multiprotease Map of the HeLa Proteome for Comprehensive Proteomics* □ S. *Mol. Cell. Proteomics* **13**, 1573–1584 (2014).

15. Garcia, B. A. *et al.* Chemical derivatization of histones for facilitated analysis by mass spectrometry. *Nat. Protoc.* **2**, 933–938 (2007).

16. Shevchenko, A., Wilm, M., Vorm, O. & Mann, M. Mass spectrometric sequencing of proteins from silver-stained polyacrylamide gels. *Anal. Chem.* **68**, 850–858 (1996).

17. Wessel, D. & Flügge, U. I. A method for the quantitative recovery of protein in dilute solution in

the presence of detergents and lipids. *Anal. Biochem.* **138**, 141–3 (1984).

18. Puchades, M., Westman, A., Blennow, K. & Davidsson, P. Removal of sodium dodecyl sulfate from protein samples prior to matrix-assisted laser desorption/ionization mass spectrometry. *Rapid Commun. Mass Spectrom.* **13**, 344–349 (1999).

19. Wiśniewski, J. R., Zougman, A., Nagaraj, N. & Mann, M. Universal sample preparation method for proteome analysis. *Nat. Methods* **6**, 359–62 (2009).

20. Fischer, R. & Kessler, B. M. Gel-aided sample preparation (GASP)-A simplified method for gel-assisted proteomic sample generation from protein extracts and intact cells. *Proteomics* **15**, 1224–1229 (2015).

21. Zougman, A., Selby, P. J. & Banks, R. E. Suspension trapping (STrap) sample preparation method for bottom-up proteomics analysis. *Proteomics* **14**, 1006–1000 (2014).

22. Hughes, C. S. *et al.* Ultrasensitive proteome analysis using paramagnetic bead technology. *Mol. Syst. Biol.* **10**, 757 (2014).

23. Batth, T. S. *et al.* Protein aggregation capture on microparticles enables multipurpose proteomics sample preparation. *Mol. Cell. Proteomics* **18**, 1027–1035 (2019).

24. Rappsilber, J., Ishihama, Y. & Mann, M. Stop and Go Extraction Tips for Matrix-Assisted Laser Desorption/Ionization, Nanoelectrospray, and LC/MS Sample Pretreatment in Proteomics. *Anal. Chem.* **75**, 663–670 (2003).

25. Neverova, I. & Van Eyk, J. E. Role of chromatographic techniques in proteomic analysis. *Journal of Chromatography B: Analytical Technologies in the Biomedical and Life Sciences* vol. 815 51–63 (2005).

26. Xie, F., Smith, R. D. & Shen, Y. Advanced proteomic liquid chromatography. *Journal of Chromatography A* vol. 1261 78–90 (2012).

27. Hsieh, E. J., Bereman, M. S., Durand, S., Valaskovic, G. A. & MacCoss, M. J. Effects of column and gradient lengths on peak capacity and peptide identification in nanoflow LC-MS/MS of complex proteomic samples. *J. Am. Soc. Mass Spectrom.* **24**, 148–153 (2013).

28. Messner, C. *et al.* Scanning SWATH acquisition enables high-throughput proteomics with chromatographic gradients as fast as 30 seconds. *bioRxiv* 656793 (2019) doi:10.1101/656793.

29. Demichev, V., Messner, C. B., Vernardis, S. I., Lilley, K. S. & Ralser, M. DIA-NN: neural networks and interference correction enable deep proteome coverage in high throughput. *Nat. Methods* **17**, 41–44 (2020).

30. Bian, Y. *et al.* Robust, reproducible and quantitative analysis of thousands of proteomes by micro-flow LC–MS/MS. *Nat. Commun.* **11**, 1–12 (2020).

31. Kelstrup, C. D. *et al.* Performance Evaluation of the Q Exactive HF-X for Shotgun Proteomics. *J. Proteome Res.* **17**, 727–738 (2018).

32. Sandow, J., Infusini, G., Dagley, L., Larsen, R. & Webb, A. Simplified high-throughput methods for deep proteome analysis on the timsTOF Pro. *bioRxiv* 657908 (2019) doi:10.1101/657908.

33. Bache, N. *et al.* A novel LC system embeds analytes in pre-formed gradients for rapid, ultra-

robust proteomics. *Mol. Cell. Proteomics* **17**, 2284–2296 (2018).

34. Wilm, M. & Mann, M. Analytical Properties of the Nanoelectrospray Ion Source. *Anal. Chem* **68**, 1–8 (1996).

35. Boesl, U. Time-of-flight mass spectrometry: Introduction to the basics. *Mass Spectrometry Reviews* vol. 36 86–109 (2017).

36. Eliuk, S. & Makarov, A. Evolution of Orbitrap Mass Spectrometry Instrumentation. *Annu. Rev. Anal. Chem.* **8**, 61–80 (2015).

37. Hunt, D. F., Yates, J. R., Shabanowitz, J., Winston, S. & Hauer, C. R. Protein sequencing by tandem mass spectrometry. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 6233–6237 (1986).

38. Steen, H. & Mann, M. The ABC's (and XYZ's) of peptide sequencing. *Nature Reviews Molecular Cell Biology* vol. 5 699–711 (2004).

39. Ting, L., Rad, R., Gygi, S. P. & Haas, W. MS3 eliminates ratio distortion in isobaric multiplexed quantitative proteomics. *Nat. Methods* **8**, 937–940 (2011).

40. Madsen, J. A., Boutz, D. R. & Brodbelt, J. S. Ultrafast ultraviolet photodissociation at 193 nm and its applicability to proteomic workflows. *J. Proteome Res.* **9**, 4205–4214 (2010).

41. Riley, N. M. & Coon, J. J. The Role of Electron Transfer Dissociation in Modern Proteomics. *Analytical Chemistry* vol. 90 40–64 (2018).

42. Robinson, M. R., Taliaferro, J. M., Dalby, K. N. & Brodbelt, J. S. 193 nm ultraviolet photodissociation mass spectrometry for phosphopeptide characterization in the positive and negative ion modes. *J. Proteome Res.* **15**, 2739–2748 (2016).

43. Kanu, A. B., Dwivedi, P., Tam, M., Matz, L. & Hill, H. H. Ion mobility-mass spectrometry. *Journal of Mass Spectrometry* vol. 43 1–22 (2008).

44. Fernandez-Lima, F., Kaplan, D. A., Suetering, J. & Park, M. A. Gas-phase separation using a trapped ion mobility spectrometer. *Int. J. Ion Mobil. Spectrom.* **14**, 93–98 (2011).

45. Meier, F. *et al.* Online parallel accumulation–serial fragmentation (PASEF) with a novel trapped ion mobility mass spectrometer. *Mol. Cell. Proteomics* **17**, 2534–2545 (2018).

46. Meier, F. *et al.* Parallel accumulation-serial fragmentation (PASEF): Multiplying sequencing speed and sensitivity by synchronized scans in a trapped ion mobility device. *J. Proteome Res.* **14**, 5378–5387 (2015).

47. Taraszka, J. A., Counterman, A. E. & Clemmer, D. E. Gas-phase separations of complex tryptic peptide mixtures. *Fresenius. J. Anal. Chem.* **369**, 234–245 (2001).

48. Meier, F. *et al.* Deep learning the collisional cross sections of the peptide universe from a million experimental values. *Nat. Commun.* **12**, 1–12 (2021).

49. Nesvizhskii, A. I. & Aebersold, R. Interpretation of Shotgun Proteomic Data. *Mol. Cell. Proteomics* **4**, 1419–1440 (2005).

50. Medzihradszky, K. F. & Chalkley, R. J. Lessons in de novo peptide sequencing by tandem mass spectrometry. *Mass Spectrometry Reviews* vol. 34 43–63 (2015).

51. Tran, N. H. *et al.* Complete de Novo Assembly of Monoclonal Antibody Sequences. *Sci. Rep.* **6**, 1–10 (2016).

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

53. 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* **20**, 3551–67 (1999).

54. Eng, J. K., McCormack, A. L. & Yates, J. R. An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* **5**, 976–89 (1994).

55. Elias, J. E. & Gygi, S. P. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214 (2007).

56. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).

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

58. Serang, O., MacCoss, M. J. & Noble, W. S. Efficient marginalization to compute protein posterior probabilities from shotgun mass spectrometry data. *J. Proteome Res.* **9**, 5346–5357 (2010).

59. Miller, R. M. *et al.* Improved Protein Inference from Multiple Protease Bottom-Up Mass Spectrometry Data. *J. Proteome Res.* **18**, 3429–3438 (2019).

60. Uszkoreit, J. *et al.* PIA: An Intuitive Protein Inference Engine with a Web-Based User Interface. *J. Proteome Res.* **14**, 2988–2997 (2015).

61. Nesvizhskii, A. I., Keller, A., Kolker, E. & Aebersold, R. A statistical model for identifying proteins by tandem mass spectrometry. *Anal. Chem.* **75**, 4646–4658 (2003).

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

63. MacLean, B. *et al.* Skyline: An open source document editor for creating and analyzing targeted proteomics experiments. *Bioinformatics* **26**, 966–968 (2010).

64. Millikin, R. J., Solntsev, S. K., Shortreed, M. R. & Smith, L. M. Ultrafast Peptide Label-Free Quantification with FlashLFQ. *J. Proteome Res.* **17**, 386–391 (2018).

65. Yu, F. *et al.* Fast Quantitative Analysis of timsTOF PASEF Data with MSFragger and IonQuant. *Mol. Cell. Proteomics* **19**, 1575–1585 (2020).

66. Cox, J. *et al.* Accurate Proteome-wide Label-free Quantification by Delayed Normalization and Maximal Peptide Ratio Extraction, Termed MaxLFQ. *Mol. Cell. Proteomics* **13**, 2513–2526 (2014).

67. Yu, F., Haynes, S. E. & Nesvizhskii, A. I. IonQuant Enables Accurate and Sensitive Label-Free Quantification With FDR-Controlled Match-Between-Runs. *Mol. Cell. Proteomics* **20**, 100077 (2021).

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

69. Monajembashi, S., Cremer, C., Cremer, T., Wolfrum, J. & Greulich, K. O. Microdissection of human chromosomes by a laser microbeam. *Exp. Cell Res.* **167**, 262–265 (1986).

70. Emmert-Buck, M. R. *et al.* Laser capture microdissection. *Science (80-.).* **274**, 998–1001 (1996).

71. Schüitze, K. & Lahr, G. Identification of expressed genes by laser-mediated manipulation of single cells. *Nat. Biotechnol.* **16**, 737–742 (1998).

72. Fustin, J. M., Karakawa, S. & Okamura, H. Circadian Profiling of Amino Acids in the SCN and Cerebral Cortex by Laser Capture Microdissection–Mass Spectrometry. *J. Biol. Rhythms* **32**, 609–620 (2017).

73. Angeles, G. *et al.* Preparative laser capture microdissection and single-pot cell wall material preparation: A novel method for tissue-specific analysis. *Planta* **224**, 228–232 (2006).

74. Hebbar, S., Schulz, W. D., Sauer, U. & Schwudke, D. Laser capture microdissection coupled with on-column extraction LC-MS n Enables lipidomics of fluorescently labeled drosophila neurons. *Anal. Chem.* **86**, 5345–5352 (2014).

75. Moulédous, L. *et al.* Proteomic analysis of immunostained, laser-capture microdissected brain samples. *Electrophoresis* **24**, 296–302 (2003).

76. Schermelleh, L. *et al.* Laser microdissection and laser pressure catapulting for the generation of chromosome-specific paint probes. *Biotechniques* **27**, 362–367 (1999).

77. Vandewoestyne, M. & Deforce, D. Laser capture microdissection in forensic research: A review. *Int. J. Legal Med.* **124**, 513–521 (2010).

78. Kerk, N. M., Ceserani, T., Lorraine Tausta, S., Sussex, I. M. & Nelson, T. M. Laser capture microdissection of cells from plant tissues. *Plant Physiol.* **132**, 27–35 (2003).

79. Amini, P. *et al.* An optimised protocol for isolation of RNA from small sections of laser-capture microdissected FFPE tissue amenable for next-generation sequencing. *BMC Mol. Biol.* **18**, 22 (2017).

80. A. Fairley, J., Gilmour, K. & Walsh, K. Making the Most of Pathological Specimens: Molecular Diagnosis in Formalin-Fixed, Paraffin Embedded Tissue. *Curr. Drug Targets* **13**, 1475–1487 (2012).

81. Uchida, Y., Sasaki, H. & Terasaki, T. Establishment and validation of highly accurate formalin-fixed paraffin-embedded quantitative proteomics by heat-compatible pressure cycling technology using phase-transfer surfactant and SWATH-MS. *Sci. Rep.* **10**, (2020).

82. Marchione, D. M. *et al.* HYPERsol: High-Quality Data from Archival FFPE Tissue for Clinical Proteomics. *J. Proteome Res.* **19**, (2020).

83. Mund, A. *et al.* AI-driven Deep Visual Proteomics defines cell identity and heterogeneity. *bioRxiv* 2021.01.25.427969 (2021) doi:10.1101/2021.01.25.427969.

84. Banks, R. E. *et al.* The potential use of laser capture microdissection to selectively obtain distinct populations of cells for proteomic analysis - Preliminary findings. in *Electrophoresis* vol. 20 689–700 (1999).

85. Emmert-Buck, M. R. *et al.* An approach to proteomic analysis of human tumors. *Mol. Carcinog.* **27**, 158–165 (2000).

86. De Marchi, T. *et al.* The advantage of laser-capture microdissection over whole tissue analysis in proteomic profiling studies. *Proteomics* **16**, 1474–1485 (2016).

87. Großerueschkamp, F. *et al.* Spatial and molecular resolution of diffuse malignant mesothelioma heterogeneity by integrating label-free FTIR imaging, laser capture microdissection and proteomics. *Sci. Rep.* **7**, (2017).

88. Asp, M. *et al.* A Spatiotemporal Organ-Wide Gene Expression and Cell Atlas of the Developing Human Heart. *Cell* **179**, 1647-1660.e19 (2019).

89. Baccin, C. *et al.* Combined single-cell and spatial transcriptomics reveal the molecular, cellular and spatial bone marrow niche organization. *Nat. Cell Biol.* **22**, 38–48 (2020).

90. Moncada, R. *et al.* Integrating microarray-based spatial transcriptomics and single-cell RNA-seq reveals tissue architecture in pancreatic ductal adenocarcinomas. *Nat. Biotechnol.* **38**, 333–342 (2020).

91. Shah, S., Lubeck, E., Zhou, W. & Cai, L. In Situ Transcription Profiling of Single Cells Reveals Spatial Organization of Cells in the Mouse Hippocampus. *Neuron* **92**, 342–357 (2016).

92. Thrane, K., Eriksson, H., Maaskola, J., Hansson, J. & Lundeberg, J. Spatially resolved transcriptomics enables dissection of genetic heterogeneity in stage III cutaneous malignant melanoma. *Cancer Res.* **78**, 5970–5979 (2018).

93. Berglund, E. *et al.* Spatial maps of prostate cancer transcriptomes reveal an unexplored landscape of heterogeneity. *Nat. Commun.* **9**, 1–13 (2018).

94. Farahbod, M. & Pavlidis, P. Untangling the effects of cellular composition on coexpression analysis. *Genome Res.* **30**, 849–859 (2020).

95. Bloom, A. B. & Zaman, M. H. Influence of the microenvironment on cell fate determination and migration. *Physiological Genomics* vol. 46 309–314 (2014).

96. De Palma, M., Biziato, D. & Petrova, T. V. Microenvironmental regulation of tumour angiogenesis. *Nature Reviews Cancer* vol. 17 457–474 (2017).

97. Balkwill, F. R., Capasso, M. & Hagemann, T. The tumor microenvironment at a glance. *J. Cell Sci.* **125**, 5591–5596 (2012).

98. Marusyk, A., Almendro, V. & Polyak, K. Intra-tumour heterogeneity: A looking glass for cancer? *Nature Reviews Cancer* vol. 12 323–334 (2012).

99. Cyll, K. *et al.* Tumour heterogeneity poses a significant challenge to cancer biomarker research. *Br. J. Cancer* **117**, 367–375 (2017).

100. Janku, F. Tumor heterogeneity in the clinic: Is it a real problem? *Therapeutic Advances in Medical Oncology* vol. 6 43–51 (2014).

101. Wang, Q. *et al.* Tumor Evolution of Glioma-Intrinsic Gene Expression Subtypes Associates with Immunological Changes in the Microenvironment. *Cancer Cell* **32**, 42-56.e6 (2017).

102. Ho, B. *et al.* Molecular subgrouping of atypical teratoid/rhabdoid tumors - A reinvestigation and

current consensus. *Neuro-Oncology* vol. 22 613–624 (2020).

103. Rudin, C. M. *et al.* Molecular subtypes of small cell lung cancer: a synthesis of human and mouse model data. *Nature Reviews Cancer* vol. 19 289–297 (2019).

104. Dienstmann, R. *et al.* Consensus molecular subtypes and the evolution of precision medicine in colorectal cancer. *Nature Reviews Cancer* vol. 17 79–92 (2017).

105. Dai, X. *et al.* Breast cancer intrinsic subtype classification, clinical use and future trends. *American Journal of Cancer Research* vol. 5 2929–2943 (2015).

106. Thomas, C. & Gustafsson, J. Å. The different roles of ER subtypes in cancer biology and therapy. *Nature Reviews Cancer* vol. 11 597–608 (2011).

107. Yates, L. R. *et al.* Subclonal diversification of primary breast cancer revealed by multiregion sequencing. *Nat. Med.* **21**, 751–759 (2015).

108. McPherson, A. *et al.* Divergent modes of clonal spread and intraperitoneal mixing in high-grade serous ovarian cancer. *Nat. Genet.* **48**, 758–767 (2016).

109. Sottoriva, A. *et al.* Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 4009–4014 (2013).

110. Stieber, D. *et al.* Glioblastomas are composed of genetically divergent clones with distinct tumourigenic potential and variable stem cell-associated phenotypes. *Acta Neuropathol.* **127**, 203–219 (2014).

111. Gerlinger, M. *et al.* Intratumor Heterogeneity and Branched Evolution Revealed by Multiregion Sequencing. *N. Engl. J. Med.* **366**, 883–892 (2012).

112. Boutros, P. C. *et al.* Spatial genomic heterogeneity within localized, multifocal prostate cancer. *Nat. Genet.* **47**, 736–745 (2015).

113. Karczewski, K. J. & Snyder, M. P. Integrative omics for health and disease. *Nat. Rev. Genet.* **19**, 299–310 (2018).

114. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nature Reviews Genetics* vol. 10 57–63 (2009).

115. Goodwin, S., McPherson, J. D. & McCombie, W. R. Coming of age: Ten years of next-generation sequencing technologies. *Nat. Rev. Genet.* **17**, 333–351 (2016).

116. Mamlouk, S. *et al.* DNA copy number changes define spatial patterns of heterogeneity in colorectal cancer. *Nat. Commun.* **8**, 14093 (2017).

117. Marek, S. *et al.* Spatial and Temporal Organization of the Individual Human Cerebellum. *Neuron* **100**, 977–993 (2018).

118. Angelo, M. *et al.* Multiplexed ion beam imaging of human breast tumors. *Nat. Med.* **20**, 436–442 (2014).

119. Giesen, C. *et al.* Highly multiplexed imaging of tumor tissues with subcellular resolution by mass cytometry. *Nat. Methods* **11**, 417–422 (2014).

120. Schwamborn, K. The Importance of Histology and Pathology in Mass Spectrometry Imaging. in

Advances in Cancer Research vol. 134 1–26 (Academic Press Inc., 2017).

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

122. Buchberger, A. R., DeLaney, K., Johnson, J. & Li, L. Mass Spectrometry Imaging: A Review of Emerging Advancements and Future Insights. *Anal. Chem.* **90**, 240–265 (2018).

123. Chughtai, K. & Heeren, R. M. A. Mass spectrometric imaging for biomedical tissue analysis. *Chem. Rev.* **110**, 3237–3277 (2010).

124. Todd, P. J., Gregory Schaaff, T., Chaurand, P. & Caprioli, R. M. Organic ion imaging of biological tissue with secondary ion mass spectrometry and matrix-assisted laser desorption/ionization. *J. Mass Spectrom.* **36**, 355–369 (2001).

125. Yang, J. & Caprioli, R. M. Matrix sublimation/recrystallization for imaging proteins by mass spectrometry at high spatial resolution. *Anal. Chem.* **83**, 5728–5734 (2011).

126. Takáts, Z., Wiseman, J. M., Gologan, B. & Cooks, R. G. Mass spectrometry sampling under ambient conditions with desorption electrospray ionization. *Science (80-.).* **306**, 471–473 (2004).

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

128. Murphy, R. C., Hankin, J. A. & Barkley, R. M. Imaging of lipid species by MALDI mass spectrometry. *Journal of Lipid Research* vol. 50 S317 (2009).

129. Baharlou, H., Canete, N. P., Cunningham, A. L., Harman, A. N. & Patrick, E. Mass Cytometry Imaging for the Study of Human Diseases—Applications and Data Analysis Strategies. *Front. Immunol.* **10**, (2019).

130. Kokkat, T. J., Patel, M. S., McGarvey, D., Livolsi, V. A. & Baloch, Z. W. Archived formalin-fixed paraffin-embedded (FFPE) blocks: A valuable underexploited resource for extraction of DNA, RNA, and protein. *Biopreserv. Biobank.* **11**, 101–106 (2013).

131. Parra, E. R., Francisco-Cruz, A. & Wistuba, I. I. State-of-the-art of profiling immune contexture in the era of multiplexed staining and digital analysis to study paraffin tumor tissues. *Cancers* vol. 11 (2019).

132. Keren, L. *et al.* A Structured Tumor-Immune Microenvironment in Triple Negative Breast Cancer Revealed by Multiplexed Ion Beam Imaging. *Cell* **174**, 1373–1387 (2018).

133. Ali, H. R. *et al.* Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat. Cancer* **1**, 163–175 (2020).

134. Kertesz, V. & Van Berkel, G. J. Fully automated liquid extraction-based surface sampling and ionization using a chip-based robotic nanoelectrospray platform. *J. Mass Spectrom.* **45**, 252–260 (2010).

135. Ryan, D. J. *et al.* Protein identification in imaging mass spectrometry through spatially targeted liquid micro-extractions. *Rapid Commun. Mass Spectrom.* **32**, 442–450 (2018).

136. Rizzo, D. G., Prentice, B. M., Moore, J. L., Norris, J. L. & Caprioli, R. M. Enhanced Spatially Resolved Proteomics Using On-Tissue Hydrogel-Mediated Protein Digestion. *Anal. Chem.* **89**,

2948–2955 (2017).

137. Guo, T. *et al.* Multi-region proteome analysis quantifies spatial heterogeneity of prostate tissue biomarkers. *Life Sci. Alliance* **1**, (2018).

138. Buczak, K. *et al.* Spatial tissue proteomics quantifies inter- and intratumor heterogeneity in hepatocellular carcinoma (HCC). *Mol. Cell. Proteomics* **17**, 810–825 (2018).

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

140. Tobler, W. R. A Computer Movie Simulating Urban Growth in the Detroit Region. *Econ. Geogr.* **46**, 234 (1970).

141. Bivand, R. S., Pebesma, E. J., Gomez-Rubio, V. & Pebesma, E. J. *Applied spatial data analysis with R*. vol. 2 (Springer, 2013).

142. Dixon, P. M. Complete Spatial Randomness. in *Wiley StatsRef: Statistics Reference Online* 1–5 (John Wiley & Sons, Ltd, 2018). doi:10.1002/9781118445112.stat08071.

143. Griffith, D. A. What is spatial autocorrelation? Reflections on the past 25 years of spatial statistics. *Espac. Geogr.* **21**, 265–280 (1992).

144. Moran, P. A. P. Notes on Continuous Stochastic Phenomena. *Biometrika* **37**, 17–23 (1950).

145. Gangodagamage, C., Zhou, X. & Lin, H. Autocorrelation, Spatial. in *Encyclopedia of GIS* 92–99 (Springer International Publishing, 2017). doi:10.1007/978-3-319-17885-1_83.

146. Stuart, T. *et al.* Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888–1902 (2019).

147. Cassese, A. *et al.* Spatial Autocorrelation in Mass Spectrometry Imaging. *Anal. Chem.* **88**, 5871–5878 (2016).

148. Xia, C., Fan, J., Emanuel, G., Hao, J. & Zhuang, X. Spatial transcriptome profiling by MERFISH reveals subcellular RNA compartmentalization and cell cycle-dependent gene expression. *Proc. Natl. Acad. Sci.* **116**, 19490–19499 (2019).

149. Nicol, P. B. Detecting Subclones from Spatially Resolved RNA-Seq Data. in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 12508 LNBI 102–107 (Springer Science and Business Media Deutschland GmbH, 2020).

150. Braz Junior, G., Cardoso de Paiva, A., Corrêa Silva, A. & Cesar Muniz de Oliveira, A. Classification of breast tissues using Moran's index and Geary's coefficient as texture signatures and SVM. *Comput. Biol. Med.* **39**, 1063–1072 (2009).

151. Abdelaal, T., Mourragui, S., Mahfouz, A. & Reinders, M. J. T. SpaGE: Spatial Gene Enhancement using scRNA-seq. *Nucleic Acids Res.* **48**, E107–E107 (2020).

152. Sun, S., Zhu, J. & Zhou, X. Statistical analysis of spatial expression patterns for spatially resolved transcriptomic studies. *Nat. Methods* **17**, 193–200 (2020).

153. Pinto Pereira, S. M., McCormack, V. A., Moss, S. M. & dos Santos Silva, I. The spatial distribution of radiodense breast tissue: A longitudinal study. *Breast Cancer Res.* **11**, 33 (2009).

154. Schmal, C., Myung, J., Herzel, H. & Bordyugov, G. Moran's I quantifies spatio-temporal pattern formation in neural imaging data. *Bioinformatics* **33**, 3072–3079 (2017).

155. Nitzan, M., Karaiskos, N., Friedman, N. & Rajewsky, N. Gene expression cartography. *Nature* **576**, 132–137 (2019).

156. Anselin, L. Local Indicators of Spatial Association—LISA. *Geogr. Anal.* **27**, 93–115 (1995).

157. Morgenstern, D. A., Gibson, S., Brown, T., Sebire, N. J. & Anderson, J. Clinical and pathological features of paediatric malignant rhabdoid tumours. *Pediatr. Blood Cancer* **54**, 29–34 (2010).

158. Biegel, J. A. *et al.* Germ-line and acquired mutations of INI1 in atypical teratoid and rhabdoid tumors. *Cancer Res.* **59**, 74–9 (1999).

159. Biegel, J. A. *et al.* Alterations of the hSNF5/INI1 gene in central nervous system atypical teratoid/rhabdoid tumors and renal and extrarenal rhabdoid tumors. *Clin. Cancer Res.* **8**, 3461–7 (2002).

160. Bookhout, C., Bouldin, T. W. & Ellison, D. W. Atypical teratoid/rhabdoid tumor with retained INI1 (SMARCB1) expression and loss of BRG1 (SMARCA4). *Neuropathology* **38**, 305–308 (2018).

161. Hasselblatt, M. *et al.* SMARCA4-mutated atypical teratoid/rhabdoid tumors are associated with inherited germline alterations and poor prognosis. *Acta Neuropathol.* **128**, 453–456 (2014).

162. Louis, D. N. *et al.* The 2016 World Health Organization Classification of Tumors of the Central Nervous System: a summary. *Acta Neuropathol.* **131**, 803–820 (2016).

163. Bikowska, B., Grajkowska, W. & Jóźwiak, J. Atypical teratoid/rhabdoid tumor: Short clinical description and insight into possible mechanism of the disease. *Eur. J. Neurol.* **18**, 813–818 (2011).

164. Richardson, E. A., Ho, B. & Huang, A. Atypical teratoid rhabdoid tumour: From tumours to therapies. *J. Korean Neurosurg. Soc.* **61**, 302–311 (2018).

165. Parwani, A. V., Stelow, E. B., Pambuccian, S. E., Burger, P. C. & Ali, S. Z. Atypical teratoid/rhabdoid tumor of the brain: Cytopathologic characteristics and differential diagnosis. *Cancer* **105**, 65–70 (2005).

166. Biswas, A., Kashyap, L., Kakkar, A., Sarkar, C. & Julka, P. K. Atypical teratoid/rhabdoid tumors: Challenges and search for solutions. *Cancer Manag. Res.* **8**, 115–125 (2016).

167. von Hoff, K. *et al.* Frequency, Risk-Factors and Survival of Children With Atypical Teratoid Rhabdoid Tumors (AT/RT) of the CNS Diagnosed between 1988 and 2004, and Registered to the German HIT Database. *Pediatr. Blood Cancer* **57**, 978–985 (2011).

168. Dufour, C. *et al.* Clinicopathologic prognostic factors in childhood atypical teratoid and rhabdoid tumor of the central nervous system: A multicenter study. *Cancer* **118**, 3812–3821 (2012).

169. Woehrer, A. *et al.* Incidence of atypical teratoid/rhabdoid tumors in children. *Cancer* **116**, 5725–5732 (2010).

170. Frühwald, M. C., Biegel, J. A., Bourdeaut, F., Roberts, C. W. M. & Chi, S. N. Atypical teratoid/rhabdoid tumors - Current concepts, advances in biology, and potential future therapies. *Neuro. Oncol.* **18**, 764–778 (2016).

171. Buscariollo, D. L., Park, H. S., Roberts, K. B. & Yu, J. B. Survival outcomes in atypical teratoid rhabdoid tumor for patients undergoing radiotherapy in a Surveillance, Epidemiology, and End Results analysis. *Cancer* **118**, 4212–4219 (2012).

172. Chi, S. N. *et al.* Intensive multimodality Treatment for children with newly diagnosed CNS atypical teratoid rhabdoid tumor. *J. Clin. Oncol.* **27**, 385–389 (2009).

173. Ginn, K. F. & Gajjar, A. Atypical teratoid rhabdoid tumor: Current therapy and future directions. *Front. Oncol.* **2 SEP**, (2012).

174. Sung, K. W. *et al.* Tandem high-dose chemotherapy and autologous stem cell transplantation for atypical teratoid/rhabdoid tumor. *Cancer Res. Treat.* **48**, 1408–1419 (2016).

175. Gardner, S. L. *et al.* Intensive induction chemotherapy followed by high dose chemotherapy with autologous hematopoietic progenitor cell rescue in young children newly diagnosed with central nervous system atypical teratoid rhabdoid tumors. *Pediatr. Blood Cancer* **51**, 235–240 (2008).

176. Finkelstein-Shechter, T. *et al.* Atypical teratoid or rhabdoid tumors: Improved outcome with high-dose chemotherapy. *J. Pediatr. Hematol. Oncol.* **32**, e182–e186 (2010).

177. Orlando, K. A., Nguyen, V., Raab, J. R., Walhart, T. & Weissman, B. E. Remodeling the cancer epigenome: mutations in the SWI/SNF complex offer new therapeutic opportunities. *Expert Rev. Anticancer Ther.* **19**, 375–391 (2019).

178. Torchia, J. *et al.* Integrated (epi)-Genomic Analyses Identify Subgroup-Specific Therapeutic Targets in CNS Rhabdoid Tumors. *Cancer Cell* **30**, 891–908 (2016).

179. Han, Z. Y. *et al.* The occurrence of intracranial rhabdoid tumours in mice depends on temporal control of Smarcb1 inactivation. *Nat. Commun.* **7**, 10421 (2016).

180. Wilson, B. G. *et al.* Epigenetic antagonism between polycomb and SWI/SNF complexes during oncogenic transformation. *Cancer Cell* **18**, 316–328 (2010).

181. Gan, L. *et al.* Epigenetic regulation of cancer progression by EZH2: From biological insights to therapeutic potential. *Biomark. Res.* **6**, 10 (2018).

182. Torchia, J. *et al.* Molecular subgroups of atypical teratoid rhabdoid tumours in children: An integrated genomic and clinicopathological analysis. *Lancet Oncol.* **16**, 569–582 (2015).

183. Johann, P. D. *et al.* Atypical Teratoid/Rhabdoid Tumors Are Comprised of Three Epigenetic Subgroups with Distinct Enhancer Landscapes. *Cancer Cell* **29**, 379–393 (2016).

184. Knipstein, J. A. *et al.* Histone deacetylase inhibition decreases proliferation and potentiates the effect of ionizing radiation in atypical teratoid/rhabdoid tumor cells. *Neuro. Oncol.* **14**, 175–183 (2012).

185. Davis, S., Scott, C., Ansorge, O. & Fischer, R. Development of a Sensitive, Scalable Method for Spatial, Cell-Type-Resolved Proteomics of the Human Brain. *J. Proteome Res.* **18**, 1787–1795 (2019).

186. Espina, V. *et al.* Laser-capture microdissection. *Nat. Protoc.* **1**, 586–603 (2006).

187. Waanders, L. F. *et al.* Quantitative proteomic analysis of single pancreatic islets. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 18902–7 (2009).

188. Thakur, D. *et al.* Microproteomic analysis of 10,000 laser captured microdissected breast tumor cells using short-range sodium dodecyl sulfate-polyacrylamide gel electrophoresis and porous layer open tubular liquid chromatography tandem mass spectrometry. *J. Chromatogr. A* **1218**, 8168–8174 (2011).

189. Clair, G. *et al.* Spatially-Resolved Proteomics: Rapid Quantitative Analysis of Laser Capture Microdissected Alveolar Tissue Samples. *Sci. Rep.* **6**, 39223 (2016).

190. Zhu, Y. *et al.* Nanodroplet processing platform for deep and quantitative proteome profiling of 10–100 mammalian cells. *Nat. Commun.* **9**, 882 (2018).

191. Zhu, Y. *et al.* Proteomic Analysis of Single Mammalian Cells Enabled by Microfluidic Nanodroplet Sample Preparation and Ultrasensitive NanoLC-MS. *Angew. Chemie Int. Ed.* **57**, 12370–12374 (2018).

192. Zhu, Y. *et al.* Subnanogram proteomics: Impact of LC column selection, MS instrumentation and data analysis strategy on proteome coverage for trace samples. *Int. J. Mass Spectrom.* **427**, 4–10 (2018).

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

194. Kano, M. & Watanabe, M. Cerebellar circuits. in *Neural Circuit and Cognitive Development (Second Edition)* (eds. Rubenstein, J., Rakic, P., Chen, B. & Kwan, K. Y.) 79–102 (Academic Press, 2020). doi:https://doi.org/10.1016/B978-0-12-814411-4.00004-4.

195. Rivara, C. B., Sherwood, C. C., Bouras, C. & Hof, P. R. Stereologic characterization and spatial distribution patterns of Betz cells in the human primary motor cortex. *Anat. Rec. - Part A Discov. Mol. Cell. Evol. Biol.* **270**, 137–151 (2003).

196. Eliuk, S. *et al.* A 'Universal' Data-dependent Mass Spectrometry Method That Eliminates Time-Consuming Method Optimization for Achieving Maximal Identifications from Each Sample. (2016).

197. Cox, J. *et al.* Andromeda: a peptide search engine integrated into the MaxQuant environment. *J. Proteome Res.* **10**, 1794–805 (2011).

198. Tyanova, S. *et al.* The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nat. Methods* **13**, 731–740 (2016).

199. R Core Team. *R: A Language and Environment for Statistical Computing*. http://www.r-project.org/ (2018).

200. Brown, G. C. Total cell protein concentration as an evolutionary constraint on the metabolic control distribution in cells. *J. Theor. Biol.* **153**, 195–203 (1991).

201. Korbo, L. & Andersen, B. B. The distributions of Purkinje cell perikaryon and nuclear volume in human and rat cerebellum with the nucleator method. *Neuroscience* **69**, 151–158 (1995).

202. Sielaff, M. *et al.* Evaluation of FASP, SP3 and iST Protocols for Proteomic Sample Preparation in the Low Microgram Range. *J. Proteome Res.* **16**, 4060–4072 (2017).

203. Dapic, I., Uwugiaren, N., Jansen, P. J. & Corthals, G. L. Fast and Simple Protocols for Mass

Spectrometry-Based Proteomics of Small Fresh Frozen Uterine Tissue Sections. *Anal. Chem.* **89**, 10769–10775 (2017).

204. Wang, H. *et al.* Development and evaluation of a micro- and nanoscale proteomic sample preparation method. *J. Proteome Res.* **4**, 2397–2403 (2005).

205. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).

206. Barski, J. J. *et al.* Calbindin in cerebellar Purkinje cells is a critical determinant of the precision of motor coordination. *J. Neurosci.* **23**, 3469–3477 (2003).

207. Sugawara, T. *et al.* Type 1 inositol trisphosphate receptor regulates cerebellar circuits by maintaining the spine morphology of purkinje cells in adult mice. *J. Neurosci.* **33**, 12186–12196 (2013).

208. Winkler, S. C., Shimobayashi, E. & Kapfhammer, J. P. PKCγ-Mediated Phosphorylation of CRMP2 Regulates Dendritic Outgrowth in Cerebellar Purkinje Cells. *Mol. Neurobiol.* **57**, (2020).

209. Brochu, G., Maler, L. & Hawkes, R. Zebrin II: A polypeptide antigen expressed selectively by purkinje cells reveals compartments in rat and fish cerebellum. *J. Comp. Neurol.* **291**, (1990).

210. Dalpé, G. *et al.* Dystonin Is Essential for Maintaining Neuronal Cytoskeleton Organization. *Mol. Cell. Neurosci.* **10**, 243–257 (1998).

211. Uhlén, M. *et al.* Tissue-based map of the human proteome. *Science (80-.).* **347**, 1260419 (2015).

212. Zhu, Y. *et al.* Spatially Resolved Proteome Mapping of Laser Capture Microdissected Tissue with Automated Sample Transfer to Nanodroplets. *Mol. Cell. Proteomics* **17**, 1864–1874 (2018).

213. Xu, K. *et al.* Benchtop-compatible sample processing workflow for proteome profiling of < 100 mammalian cells. *Anal. Bioanal. Chem.* **411**, 4587–4596 (2019).

214. Petralia, F. *et al.* Integrated Proteogenomic Characterization across Major Histological Types of Pediatric Brain Cancer. *Cell* **183**, 1962–1985 (2020).

215. Tangrea, M. A. *et al.* Effect of immunohistochemistry on molecular analysis of tissue samples: Implications for microdissection technologies. *J. Histochem. Cytochem.* **59**, 591–600 (2011).

216. Bivand, R. S. & Wong, D. W. S. Comparing implementations of global and local indicators of spatial association. *TEST* **27**, 716–748 (2018).

217. Hijmans, R. J. raster: Geographic Data Analysis and Modeling. (2020).

218. Yu, G. *et al.* GOSemSim: an R package for measuring semantic similarity among GO terms and gene products. *Bioinformatics* **26**, 976–978 (2010).

219. Yu, G., Wang, L.-G., Han, Y. & He, Q.-Y. clusterProfiler: an R Package for Comparing Biological Themes Among Gene Clusters. *Omi. A J. Integr. Biol.* **16**, 284–287 (2012).

220. Carlson, M. org.Hs.eg.db: Genome wide annotation for Human. (2020).

221. Bankhead, P. *et al.* QuPath: Open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).

222. Ludwig, K. R., Schroll, M. M. & Hummon, A. B. Comparison of In-Solution, FASP, and S-Trap Based Digestion Methods for Bottom-Up Proteomic Studies. *J. Proteome Res.* **17**, 2480–2490 (2018).

223. Porlier, M. M., de Néchaud, B. & Gros, F. Peripherin, a new member of the intermediate filament protein family. *Dev. Neurosci.* **6**, 335–344 (1983).

224. Yuan, A. *et al.* Peripherin is a subunit of peripheral nerve neurofilaments: Implications for differential vulnerability of cns and peripheral nervous system axons. *J. Neurosci.* **32**, 8501–8508 (2012).

225. Hol, E. M. & Capetanaki, Y. Type III intermediate filaments desmin, glial fibrillary acidic protein (GFAP), vimentin, and peripherin. *Cold Spring Harbor Perspectives in Biology* vol. 9 a021642 (2017).

226. Eriksson, K. S. *et al.* The type III neurofilament peripherin is expressed in the tuberomammillary neurons of the mouse. *BMC Neurosci.* **9**, 26 (2008).

227. Xiao, S. *et al.* An aggregate-inducing peripherin isoform generated through intron retention is upregulated in amyotrophic lateral sclerosis and associated with disease pathology. *J. Neurosci.* **28**, 1833–1840 (2008).

228. Troy, C. M., Brown, K., Greene, L. A. & Shelanski, M. L. Ontogeny of the neuronal intermediate filament protein, peripherin, in the mouse embryo. *Neuroscience* **36**, 217–237 (1990).

229. Wiśniewski, J. R., Hein, M. Y., Cox, J. & Mann, M. A "proteomic ruler" for protein copy number and concentration estimation without spike-in standards. *Mol. Cell. Proteomics* **13**, 3497–506 (2014).

230. Alberts, B. *et al. Molecular Biology of the Cell*. (Garland, 2002).

231. Gröbner, S. N. *et al.* The landscape of genomic alterations across childhood cancers. *Nature* **555**, 321–327 (2018).

232. Sickmann, H. M., Walls, A. B., Schousboe, A., Bouman, S. D. & Waagepetersen, H. S. Functional significance of brain glycogen in sustaining glutamatergic neurotransmission. *J. Neurochem.* **109**, 80–86 (2009).

233. Xu, J. *et al.* Requirement of glycogenolysis for uptake of increased extracellular K + in astrocytes: Potential implications for K+ homeostasis and glycogen usage in brain. *Neurochem. Res.* **38**, 472–485 (2013).

234. Duran, J., Saez, I., Gruart, A., Guinovart, J. J. & Delgado-García, J. M. Impairment in long-term memory formation and learning-dependent synaptic plasticity in mice lacking glycogen synthase in the brain. *J. Cereb. Blood Flow Metab.* **33**, 550–556 (2013).

235. Suzuki, A. *et al.* Astrocyte-Neuron Lactate Transport Is Required for Long-Term Memory Formation. *Cell* **144**, 810–823 (2011).

236. Bak, L. K. & Walls, A. B. Astrocytic glycogen metabolism in the healthy and diseased brain. *J. Biol. Chem.* **293**, 7108–7116 (2018).

237. Brown, A. M. & Ransom, B. R. Astrocyte glycogen and brain energy metabolism. *Glia* **55**, 1263–1271 (2007).

238. Saez, I. *et al.* Neurons have an active glycogen metabolism that contributes to tolerance to hypoxia. *J. Cereb. Blood Flow Metab.* **34**, 945–955 (2014).

239. Favaro, E. *et al.* Glucose Utilization via Glycogen Phosphorylase Sustains Proliferation and Prevents Premature Senescence in Cancer Cells. *Cell Metab.* **16**, 751–764 (2012).

240. Newgard, C. B., Hwang, P. K. & Fletterick, R. J. The Family of Glycogen Phosphorylases: Structure and Functio. *Crit. Rev. Biochem. Mol. Biol.* **24**, 69–99 (1989).

241. Mathieu, C. *et al.* An Isozyme-specific Redox Switch in Human Brain Glycogen Phosphorylase Modulates Its Allosteric Activation by AMP. *J. Biol. Chem.* **291**, 23842–23853 (2016).

242. Tafani, M. *et al.* The Interplay of Reactive Oxygen Species, Hypoxia, Inflammation, and Sirtuins in Cancer Initiation and Progression. *Oxid. Med. Cell. Longev.* **2016**, 1–18 (2016).

243. Winter, S. C. *et al.* Relation of a Hypoxia Metagene Derived from Head and Neck Cancer to Prognosis of Multiple Cancers. *Cancer Res.* **67**, 3441–3449 (2007).

244. Caramel, J., Quignon, F. & Delattre, O. RhoA-Dependent Regulation of Cell Migration by the Tumor Suppressor hSNF5/INI1. *Cancer Res.* **68**, 6154–6161 (2008).

245. Yanagida, K. *et al.* Size-selective opening of the blood-brain barrier by targeting endothelial sphingosine 1-phosphate receptor 1. *Proc. Natl. Acad. Sci. U. S. A.* **114**, 4531–4536 (2017).

246. Zhu, D. *et al.* Protein S controls hypoxic/ischemic blood-brain barrier disruption through the TAM receptor Tyro3 and sphingosine 1-phosphate receptor. *Blood* **115**, 4963–4972 (2010).

247. Wallingford, N. *et al.* Prolylcarboxypeptidase regulates food intake by inactivating α-MSH in rodents. *J. Clin. Invest.* **119**, 2291–2303 (2009).

248. Palmiter, R. D. Reduced levels of neurotransmitter-degrading enzyme PRCP promote obesity. *J. Clin. Invest.* **119**, 2130–2133 (2009).

249. Vaillant, C. *et al.* Protease nexin 1 and its receptor LRP modulate SHH signalling during cerebellar development. *Development* **134**, 1745–1754 (2007).

250. Brunner, A.-D. *et al.* Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *bioRxiv* 2020.12.22.423933 (2020) doi:10.1101/2020.12.22.423933.

251. Webb-Robertson, B. J. M. *et al.* Review, evaluation, and discussion of the challenges of missing value imputation for mass spectrometry-based label-free global proteomics. *J. Proteome Res.* **14**, 1993–2001 (2015).

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

253. Bramer, L. M., Irvahn, J., Piehowski, P. D., Rodland, K. D. & Webb-Robertson, B. J. M. A Review of Imputation Strategies for Isobaric Labeling-Based Shotgun Proteomics. *J. Proteome Res.* **20**, 1–13 (2021).

254. Medo, M., Aebersold, D. M. & Medová, M. ProtRank: Bypassing the imputation of missing values in differential expression analysis of proteomic data. *BMC Bioinformatics* **20**, 1–12 (2019).

255. Jin, L. *et al.* A comparative study of evaluating missing value imputation methods in label-free

proteomics. *Sci. Rep.* **11**, 1760 (2021).

256. Li, Z., Song, T., Yong, J. & Kuang, R. Imputation of spatially-resolved transcriptomes by graph-regularized tensor completion. *PLoS Comput. Biol.* **17**, e1008218 (2021).

257. Langfelder, P., Zhang, B. & Horvath, S. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R. *Bioinformatics* **24**, 719–720 (2008).

258. Krug, K. *et al.* A Curated Resource for Phosphosite-specific Signature Analysis. *Mol. Cell. Proteomics* **18**, 576–593 (2019).

259. Liberzon, A. *et al.* Molecular signatures database (MSigDB) 3.0. *Bioinformatics* **27**, 1739–1740 (2011).

260. Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108–112 (2009).

261. Deutskens, F., Yang, J. & Caprioli, R. M. High spatial resolution imaging mass spectrometry and classical histology on a single tissue section. *J. Mass Spectrom.* **46**, 568–571 (2011).

262. Ly, A. *et al.* Site-to-Site Reproducibility and Spatial Resolution in MALDI–MSI of Peptides from Formalin-Fixed Paraffin-Embedded Samples. *Proteomics - Clin. Appl.* **13**, 1800029 (2019).

263. Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38–47 (2019).

264. Lima, L. G. & Monteiro, R. Q. Activation of blood coagulation in cancer: Implications for tumour progression. *Biosci. Rep.* **33**, 701–710 (2013).

265. Mukaihara, K. *et al.* Expression of F-actin-capping protein subunit beta, CAPZB, is associated with cell growth and motility in epithelioid sarcoma. *BMC Cancer* **16**, 1–13 (2016).

266. Hay, S. B., Ferchen, K., Chetal, K., Grimes, H. L. & Salomonis, N. The Human Cell Atlas bone marrow single-cell interactive web portal. *Exp. Hematol.* **68**, 51–61 (2018).

267. Patel, A. P. *et al.* Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma. *Science (80-.).* **344**, 1396–1401 (2014).

268. Darmanis, S. *et al.* Single-Cell RNA-Seq Analysis of Infiltrating Neoplastic Cells at the Migrating Front of Human Glioblastoma. *Cell Rep.* **21**, 1399–1410 (2017).

269. Puchalski, R. B. *et al.* An anatomic transcriptional atlas of human glioblastoma. *Science (80-.).* **360**, 660–663 (2018).

270. Kouzarides, T. Chromatin Modifications and Their Function. *Cell* **128**, 693–705 (2007).

271. Erkek, S. *et al.* Comprehensive Analysis of Chromatin States in Atypical Teratoid/Rhabdoid Tumor Identifies Diverging Roles for SWI/SNF and Polycomb in Gene Regulation. *Cancer Cell* **35**, 95–110 (2019).

272. Hasselblatt, M., Johann, P. D., Kool, M. & Frühwald, M. C. Reduced histone H3 K27 trimethylation is encountered in about 50% of atypical teratoid/rhabdoid tumors (AT/RT) but is not associated with molecular subgroup status and outcome. *Acta Neuropathol.* **134**, 817–818 (2017).

273. Maier, S. K. *et al.* Comprehensive identification of proteins from MALDI imaging. *Mol. Cell.*

Proteomics **12**, 2901–2910 (2013).

274. St John, E. R. *et al.* Rapid evaporative ionisation mass spectrometry of electrosurgical vapours for the identification of breast pathology: Towards an intelligent knife for breast cancer surgery. *Breast Cancer Res.* **19**, 1–14 (2017).

275. Holzlechner, M., Eugenin, E. & Prideaux, B. Mass spectrometry imaging to detect lipid biomarkers and disease signatures in cancer. *Cancer Rep.* **2**, (2019).

276. Kanwal, M., Smahel, M., Olsen, M., Smahelova, J. & Tachezy, R. Aspartate ß-hydroxylase as a target for cancer therapy. *J. Exp. Clin. Cancer Res.* **39**, 1–12 (2020).

277. Lawton, M., Tong, M., Gundogan, F., Wands, J. R. & De La Monte, S. M. Aspartyl-(asparaginyl) β-hydroxylase, hypoxia-inducible factor-1α and Notch cross-talk in regulating neuronal motility. *Oxid. Med. Cell. Longev.* **3**, 347–356 (2010).

278. Hadas, R. *et al.* Hyaluronan control of the primary vascular barrier during early mouse pregnancy is mediated by uterine NK cells. *JCI Insight* **5**, (2020).

279. Kim, M. *et al.* VEGF-A regulated by progesterone governs uterine angiogenesis and vascular remodelling during pregnancy. *EMBO Mol. Med.* **5**, 1415–1430 (2013).

280. Elosua-Bayes, M., Nieto, P., Mereu, E., Gut, I. & Heyn, H. SPOTlight: seeded NMF regression to deconvolute spatial transcriptomics spots with single-cell transcriptomes. *Nucleic Acids Res.* **49**, e50–e50 (2021).

281. Andersson, A. *et al.* Single-cell and spatial transcriptomics enables probabilistic inference of cell type topography. *Commun. Biol.* **3**, 1–8 (2020).

282. Dong, R. & Yuan, G. C. SpatialDWLS: accurate deconvolution of spatial transcriptomic data. *Genome Biol.* **22**, 1–10 (2021).

283. Wilhelm, M. *et al.* Mass-spectrometry-based draft of the human proteome. *Nature* **509**, 582–587 (2014).

284. Hartlmayr, D. *et al.* An automated workflow for label-free and multiplexed single cell proteomics sample preparation at unprecedented sensitivity. *bioRxiv* 2021.04.14.439828 (2021) doi:10.1101/2021.04.14.439828.

285. Woo, J. *et al.* High-throughput and high-efficiency sample preparation for single-cell 1 proteomics using a nested nanowell chip 2. *bioRxiv* 2021.02.17.431689 (2021) doi:10.1101/2021.02.17.431689.

286. Petelski, A. A. *et al.* Multiplexed single-cell proteomics using SCoPE2. *bioRxiv* 2021.03.12.435034 (2021) doi:10.1101/2021.03.12.435034.

287. Boluki, S., Zamani Dadaneh, S., Qian, X. & Dougherty, E. R. Optimal clustering with missing values. *BMC Bioinformatics* **20**, 1–10 (2019).

288. Meier, F. *et al.* diaPASEF: parallel accumulation–serial fragmentation combined with data-independent acquisition. *Nat. Methods* **17**, 1229–1236 (2020).

289. Santos, A. *et al.* Clinical Knowledge Graph Integrates Proteomics Data into Clinical Decision-Making. *bioRxiv* 2020.05.09.084897 (2020) doi:10.1101/2020.05.09.084897.

290. Subramanian, I., Verma, S., Kumar, S., Jere, A. & Anamika, K. Multi-omics Data Integration, Interpretation, and Its Application. *Bioinform. Biol. Insights* **14**, (2020).

291. Quanico, J. *et al.* Development of liquid microjunction extraction strategy for improving protein identification from tissue sections. *J. Proteomics* **79**, 200–218 (2013).

292. Drummond, E. S., Nayak, S., Ueberheide, B. & Wisniewski, T. Proteomic analysis of neurons microdissected from formalin-fixed, paraffin-embedded Alzheimer's disease brain tissue. *Sci. Rep.* **5**, 15456 (2015).

293. Molina, M. *et al.* Enrichment of single neurons and defined brain regions from human brain tissue samples for subsequent proteome analysis. *J. Neural Transm.* **122**, 993–1005 (2015).

294. Wisztorski, M. *et al.* Spatially-resolved protein surface microsampling from tissue sections using liquid extraction surface analysis. *Proteomics* **16**, 1622–1632 (2016).

295. Donnarumma, F. & Murray, K. K. Laser ablation sample transfer for localized LC-MS/MS proteomic analysis of tissue. *J. Mass Spectrom.* **51**, 261–268 (2016).

296. Quanico, J., Franck, J., Wisztorski, M., Salzet, M. & Fournier, I. Integrated mass spectrometry imaging and omics workflows on the same tissue section using grid-aided, parafilm-assisted microdissection. *Biochim. Biophys. Acta - Gen. Subj.* **1861**, 1702–1714 (2017).

297. Dilillo, M. *et al.* Mass Spectrometry Imaging, Microdissection and LC-MS/MS of the Same Tissue Section. *J. Proteome Res.* **16**, acs.jproteome.7b00284 (2017).

298. Djuric, U. *et al.* Spatiotemporal proteomic profiling of human cerebral development. *Mol. Cell. Proteomics* **16**, 1558–1562 (2017).

299. García-Berrocoso, T. *et al.* Single cell immuno-laser microdissection coupled to label-free proteomics to reveal the proteotypes of human brain cells after ischemia. *Mol. Cell. Proteomics* **17**, 175–189 (2018).