Analysis of Cellular Heterogeneity in Breast Cancer by Single Cell Sequencing

Dr. Karen Sayal

St. Cross College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Trinity 2022

Abstract

Breast cancer is a complex heterogenous 3D ecosystem. The heterogenous composition of breast cancer determines disease progression and treatment responses. Triple receptor negative breast cancer (TNBC) is a distinct subtype with poor clinical outcomes. Deconvolution of spatially-regulated transcriptomic and microenvironmental drivers unique to TNBC offers the potential to reveal new therapeutic vulnerabilities.

Single cell RNA sequencing (scRNA-seq) and spatial transcriptomic technologies were applied to three treatment naive patient-derived breast cancer samples.

New spatial transcriptomic and scRNA-seq experimental pipelines were established. The new technologies were successfully applied to clinical grade biopsy samples. Cellular heterogeneity within the epithelial and non-epithelial compartment was identified across the three samples. The heterogeneity identified is consistent with the published literature.

Knowledge in the theoretical underpinnings for scRNA-seq analysis along with the skills required for data analysis in a small patient cohort were acquired during the DPhil. The application of algebraic topology, manifold learning and graph theory in evaluating and interpreting scRNA-seq has been studied.

The computational tools available for integrating spatial transcriptomics and scRNA-seq data were critically appraised. Future perspectives on approaches for multimodal integration were explored.

Acknowledgements

The thesis is the culmination of an enormous amount of inter-disciplinary work and close collaboration with many scientists spread across several countries.

I wish to thank my supervisors Professors Francesca Buffa and Adrian Harris for their guidance. I appreciate the help from all members of the Buffa and Harris lab.

I am grateful for the invaluable support and guidance provided by collaborators including Miss Ashvina Segaran, Mr. Thomas Carroll, Dr. Sophie Kirschner and Dr. Stan Ng at the Ludwig Institute of Cancer Research, Professor Evan Macosko and Dr. Evan Murray at the Broad Institute of MIT and Harvard, Dr. Morgane Rouault at CARTANA in Sweden, Mr. Connor Scott in the Nuffield Department of Clinical Neurosciences, Mr. Simon Davies in the Target Discovery Institute, Dr. Stephanie Jones in the Oxford Radcliffe Biobank, Dr. Simon Lord and Dr. Alistair Easton in the Department of Oncology and Dr. Neil Ashley at the WIMM Single Cell Facility.

I wish to acknowledge the valuable support and help of the Cancer Research UK Oxford Centre for generously supporting my Clinical Research Training Fellowship without whom this huge amount of work could not have been completed.

Declaration of Authorship

I confirm that the work submitted in this thesis is wholly my own work unless otherwise indicated.

Contents

Abstract **iii**

Acknowledgements **v**

List of Figures **xi**

List of Tables **xv**

List of Tables **xvi**

1 Introduction **1**

1.1 The Cellular Architecture of the Breast 2

1.1.1 Macroscopic Architecture 2

1.1.2 Microscopic Architecture 2

1.2 Triple Negative Breast Cancer 6

1.2.1 Low-Grade TNBC 7

1.2.2 High-Grade TNBC 7

1.3 The Molecular and Spatial Landscape of Breast Cancer 14

1.3.1 Molecular Landscape 14

1.3.2 Spatial Landscape 16

1.4 Intention of the DPhil Project 23

2 Experimental Design **25**

2.1 Technologies Available at the Outset of the DPhil 26

2.2 Intended Technology Deployment 26

2.3 Critical Analysis of Spatial Transcriptomic Techniques 27

2.3.1 Overview of Techniques 27

2.3.2 Selection of Techniques 30

2.4 Protocols 31

2.5 Pandemic-related DPhil Disruption 36

3 Application of Spatial-omics Technologies in Breast Cancer 37
3.1 Opportunities of Spatial Biology in Cancer Medicine 39
3.1.1 The Application of Spatial-omics 39
3.1.2 Insights offered by Spatial-omics 46
3.1.3 Opportunities in Data Interpretation 57
3.1.4 Computational Tools 58
3.2 Methods and Results 62
3.2.1 Experimental Pipelines for Spatial Sequencing 62
3.2.2 Slide-seq: Experimental Pipeline 63
3.2.3 Slide-seq: Optimisation for Solid Tumours 64
3.2.4 Slide-seq: Application to Human Tissue 68
3.2.5 Slide-seq: Results 72
3.2.6 CARTANA: Commercial Pipeline 75
3.2.7 CARTANA: Bespoke Spatial Hypoxia Panel 77
3.2.8 CARTANA: Results 78
3.3 Challenges and Limitations 82
3.3.1 Quality Control 82
3.3.2 New Computational Tools 87
3.3.3 Limitations 89

4 A Single Cell Atlas of Breast Cancer 91
4.1 Introduction 92
4.2 Experimental and Bioinformatics Pipelines 92
4.2.1 Sample Collection 92
4.3 Results 93
4.3.1 Patient Specific Cell and Gene QC Thresholds 94
4.3.2 Critical Analysis of the Machine Learning Tools used in Single Cell Biology 97
4.3.3 Canonical Lineage Markers for Cell Type Assignment . . . 104
4.3.4 Future Considerations 114
4.4 Discussion 120
4.4.1 Limitations 121
4.4.2 Future Directions 121

5 Multimodal Profiling of Breast Cancer 123
5.1 Multimodal Integration 124
5.1.1 Definition 124
5.1.2 Benefit 127
5.1.3 Challenges 128
5.2 Current Methods 129

5.2.1 Weighted Nearest Neighbour 129
5.2.2 MultiMap 134
5.2.3 Multi-Omics Factor Analysis v2 139
5.3 Future Methods 145
5.3.1 Multimodal Variational Autoencoders 146
5.3.2 Diffusion Models 148
5.4 Conclusion 159

6 Conclusion **161**

Bibliography **165**

Appendices **188**

A **191**

List of Figures

1.1 Heterogeneity of the mammary luminal epithelial subpopulation 5
1.2 Pseduotemporal differentiation trajectory of the mammary epithelium 5
1.3 Relationship between TNBC and Intrinsic subtypes 8
1.4 TNBC samples exhibit a combination of molecular subtypes . . . 12
1.5 The malignant epithelial subpopulation exhibits transcriptional heterogeneity 13
1.6 KM survival curves for TNBC patients from the METABRIC cohort 13
1.7 IntClust classification correlates with (a) survival outcomes and (b) chemotherapy sensitivity 15
1.8 Epithelial and TME cell phenotypes 17
1.9 Comparison of Shannon entropy between epithelial and TME phenotypes across breast cancer subtypes 19
1.10 Cell types enriched at the tumour-stroma interface 20
1.11 Cell types enriched at the perivascular interface 20
1.12 Multicellular TME structures 21

3.1 BaSISS Workflow 41
3.2 Spatial Clone Maps in Breast Cancer 42
3.4 Spatial Subclonal Structure in Nodal Metastatic Disease 45
3.5 Biological Insights of Spatial-omics 46
3.6 Heatmap comparing Ki-67 Positivity and MPI Score 48
3.7 Spatial Maps of MPI Score and Correlation Metrics 49
3.8 ccD-CMD and Cell Cycle Coherence Metrics 50
3.9 Cell Cycle Coherence in HER2 Positive Breast Cancer 50
3.10 MPI Maps can be linked with Histology 51
3.11 MPI Maps can be linked with Histology 53
3.12 Architecture of the Vision Transformer 61
3.13 H&E Assessment of Breast Cancer Xenografts 64
3.14 Slide-seq Sample Preparation 65
3.15 Sample Orientation during Slide-seq Tissue Sectioning 65
3.16 Slide-seq Protocol 66

3.17 Optimisation of RNA Hybridisation for Slide-seq 67
3.18 H&E Assessment of Clinical Breast Cancer Biopsies 70
3.19 Desirable Histological Features 71
3.20 Less Desirable Histological Features 72
3.21 UMI Spatial Map of Breast Cancer in Slide-seq 73
3.22 Cell type spatial map of normal and malignant prostate tissue . . 74
3.23 Histological Assessment of ORB Samples 76
3.24 Single Cell Expression of Hypoxia-Associated Genes 78
3.25 CARTANA Spatial Expression Profile of Hypoxia Markers 79
3.26 Cell Segmentation using the Watershed Transform Algorithm . . 80
3.27 Deep-Learning Cell Segmentation 81
3.28 Overview of the Tangram Algorithm 88
3.29 Optimisation Objective of the Tangram Algorithm 88

4.1 Sample BC18 Filtering Thresholds 96
4.2 Viable Single Cell Selection 97
4.3 Scree Plot 99
4.4 Principal Component Selection 99
4.5 Cross Entropy Loss Function 104
4.6 Single Cell 2D UMAP Projections 105
4.7 Manual Single Cell Annotation 109
4.7 Manual Single Cell Annotation 110
4.8 Single Cell Data Integration with Harmony 111
4.9 Comparison of the UMAP Embeddings of Normal and Malignant Mammary Tissue 113
4.10 2D UMAP Embeddings can introduce Data Distortions 116
4.11 Picasso Embeddings of Ex Utero Mouse Embryo scRNA-seq Data 117
4.12 Correlation Benchmarks for Ex Utero Mouse Embyro scRNA-seq Data 117
4.13 MCL Autoencoder Architecture 118
4.14 Differentially Expressed Genes 119

5.1 The directional classification system for multimodal integration . 126
5.2 WNN Analysis 131
5.3 MultiMap Algorithm 135
5.4 Multimap Integration of Spatial Transcriptomics and Gene Expression Data 136
5.5 Structure of data input for MOFA+ 140
5.6 Graphical model for MOFA+ multimodal matrix factorisation . . 141
5.7 MMVAE generates meaningful cross-modality reconstructions . 148

5.8 Correlation values for MMVAE 148
5.9 Directed Graphical Model for Forward and Reverse Diffusion Processes 149
5.10 Diffusion generated image of the gross pathology of colorectal cancer 153
5.11 Diffusion generated image of the microscopic pathology of colorectal cancer 154
5.12 Diffusion generated image of the microscopic pathology of hepatocellular cancer 155
5.13 Diffusion generated image of tumour-immune dynamics in colorectal cancer 156
5.14 Diffusion generated image of the spatial transcriptomic subclonal architecture in breast cancer 157

List of Tables

1.1 Summary of TNBC Molecular Subtypes [13] 10
1.2 Spatially Preserved TME Structures [23] 21

2.1 Comparison of Spatial Transcriptomic Technologies 30
2.2 Slide-seq Custom Primers 32

3.1 MPI Scoring System. Reproduced from [43]. 47
3.2 Summary of ST Computational Tools 59
3.3 Cell Type Proportions per Breast Cancer Biopsy 71

4.1 Post-Surgical Pathological Characteristics of Breast Cancer Samples 93
4.2 Thresholds for Total Counts 95
4.3 Thresholds for Total Features 95
4.4 Thresholds for Mitochondrial Counts 95
4.5 Canonical Markers for Non-Immune Cell Type Assignment 107
4.6 Canonical Markers for Adaptive Immune Cell Type Assignment . 107
4.7 Canonical Markers for Innate Immune Cell Type Assignment . . . 107

5.1 Text for Diffusion-Based Generative Imaging 151

A.1 HUGO gene names for the CARTANA hypoxia-immune bespoke panel 192

List of Tables

1 Introduction

Contents

1.1 The Cellular Architecture of the Breast 2
1.1.1 Macroscopic Architecture 2
1.1.2 Microscopic Architecture 2
1.2 Triple Negative Breast Cancer 6
1.2.1 Low-Grade TNBC 7
1.2.2 High-Grade TNBC 7
1.3 The Molecular and Spatial Landscape of Breast Cancer 14
1.3.1 Molecular Landscape 14
1.3.2 Spatial Landscape 16
1.4 Intention of the DPhil Project 23

1.1 The Cellular Architecture of the Breast

The breast consists of three sub-compartments: the ductal-lobular network, the stroma and the nipple. Each sub-compartment can be considered at two spatial levels: macroscopic and microscopic.

The macroscopic architecture encompasses the gross pathological definition of breast tissue. The microscopic architecture arises from a combination of well-established histological perspectives together with FACS-sorting and gene expression analysis of normal breast tissue.

1.1.1 Macroscopic Architecture

The breast consists of mammary glands embedded within fibroadipose tissue which converge at the nipple [1].

The mammary glands are composed of lobules and ducts. The lobules are the site of milk production and have an acinar configuration. Milk produced in the lobules are transported in a complex network of ducts which ultimately drain to the nipple.

The mammary glands are supported by the suspensory ligaments of Cooper. The ligaments attach the mammary tissue to the overlying dermis and underlying pectoral fascia. They also function to separate the mammary lobules. The compartment between the mammary glands and fibrous tissue consists of adipose tissue.

1.1.2 Microscopic Architecture

1.1.2.1 Histological Structure of the Breast

The lobules are modified sweat glands [1]. The network of ducts which drain via lactiferous sinuses at the nipple are bilayered consisting of:

1. Inner luminal epithelial cells
2. Outer basal myoepithelium

Luminal cells are cuboidal to columnar epithelium [1]. The myoepithelium is varied, ranging from flat cells to prominent epithlioid cells . The basement membrane is composed of type IV collagen and laminin. It surrounds the ducts and lobules and separates the ductal-lobular system from the stroma.

The stroma consists of fibroadipose tissue [1]. Its composition varies according to spatial location. Intralobular stroma is dense with high collagen content. In contrast, interlobular stroma contains inflammatory cells with less collagen.

The central nipple is composed of dense stroma, smooth muscle and multiple sebaceous glands [1].

1.1.2.2 Physiological States of the Breast

Breast tissue exhibits characteristic histological changes during the menstrual cycle, pregnancy and menopause.

In the follicular phase of the menstrual cycle, the breast is largely quiescent with small lobules and myoepithelial cells and minimal inflammatory infiltrate [1]. In the luteal phase, the lobules increase with stromal oedema (accounting for breast fullness) and myoepithelial cells develop marked cytoplasmic vacuolisation [1]. Stromal immune infiltration increases.

During pregnancy, there is a marked increase in the number and size of lobules accompanied by a decrease in the stromal compartment [1]. In lactation, the epithelial cells develop a bulbous morphology with an increase in secretory material [1].

In menopause, there is atrophy and loss of architecture of the lobules with a decrease in volume of intralobular stroma [1]. The stroma becomes hyalinzed. Glandular tissue becomes generally replaced by adipose tissue.

1.1.2.3 Differentiation Hierarchy of Normal Mammary Tissue

Advances offered by FACS-sorting, survival assays and scRNA-seq have given further insight into the traditional histological characterisation of normal breast tissue and shed light on the intrinsic structural complexity conferring its regenerative capacity in distinct physiological states.

Earlier work using FACS-sorting and survival assays identified mature ER^+ luminal, basal and luminal progenitor (LP) cells [2]. LP cells are cells with restricted differentiation capacity and arise from phenotypically distinct mammary stem cells, a rare cell sub-population enriched within the basal layer [3]. Further heterogeneity was identified in the LP population with $ALDH^+$, $ALDH^-$ and $ERBB3^-$ progenitors [2]. $ALDH^+$ progenitors were found to have a gene signature similar to basal-like breast cancer [2].

More recent work using scRNA-seq has demonstrated similar heterogeneity in the luminal epithelial compartment with the existence of basal, myoepithelial and two luminal subpopulations, secretory L1 cells and hormone-responsive L2 cells (Figure 1.1) [4]. L1 cells correlate closely with luminal progenitor cells with potential capacity for alveologenesis and L2 cells with mature luminal cells [4]. Basal cells are enriched for expression of stem cell markers, such as TCF4 [4]. Myoepithelial cells are associated with markers linked to integrin and paxillin signalling for maintenance of the physical architecture of the breast [4].

Spatial localisation of the identified cell types generally correlated with the known histological architecture of the breast. Luminal markers, eg KRT8 and basal markers, eg KRT14 were most strongly expressed in luminal and basal cells respectively [4]. However, exceptions were identified. Several regions of lobular tissue showed $KRT8^+/KRT14^+$ expression patterns [4]. It is the first time a dual expression pattern has been identified in human mammary tissue.

Pseudotemporal ordering of the cell states suggests a continuous single

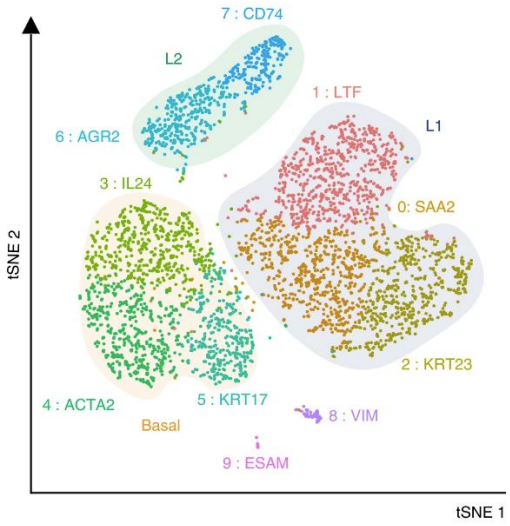


Figure 1.1: Heterogeneity of the mammary luminal epithelial subpopulation. Reproduced from [4]. Three cell types were identified within ten clusters: L1, L2 and basal. L1 cells are secretory luminal cells and L2 cells are hormone-responsive luminal cells. Marker genes associated with each cluster are shown. Marker genes, 0-9, are colour-coded.

trajectory in which L1, L2 and myoepithelial states are connected through a common basal state, a state known to be enriched for stem cell markers (Figure 1.2) [4]. This differentiation pathway is consistent with previous models of mammary differentiation.

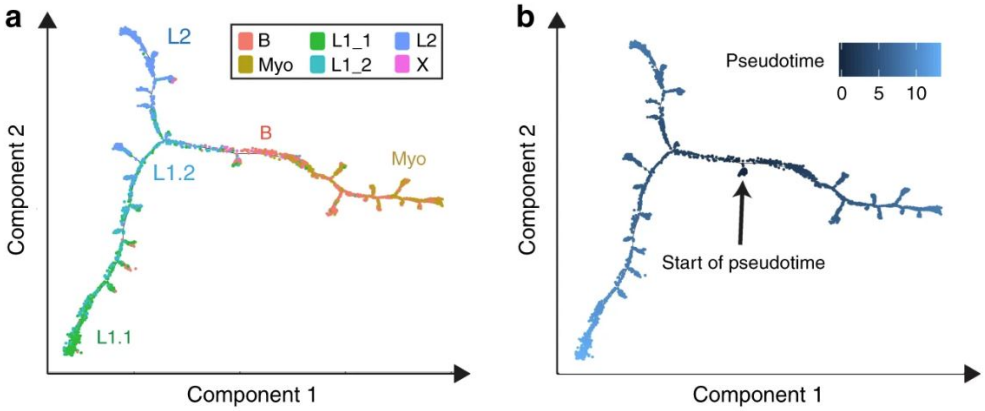


Figure 1.2: Pseduotemporal differentiation trajectory of the mammary epithelium. Reproduced from [4].

By comparing cell type gene signatures with signatures linked with breast cancer subtypes, Nguygen et al inferred the relationship between homeostatic mammary tissue and breast cancer subtypes. L2 luminal cells are closely linked with Luminal A and B breast cancer, myoepithelial cells with mesenchymal-like TNBC subtype and L1 cells with basal-like TNBC [4]. It supports the model that different breast cancer subtypes arise from distinct cells-of-origin.

1.2 Triple Negative Breast Cancer

Triple negative breast cancer (TNBC), as defined by lack of expression of ER, PR and HER2, constitutes approximately 20% of all breast cancers [5]. It is typically more clinically and biologically aggressive than other breast cancer subtypes [6]. It is more likely to present in younger patients with larger size, nodal disease at initial presentation and with a greater propensity to develop metastases [6].

The current mainstay of treatment is chemotherapy. TNBC exhibits high response rates to chemotherapy. However, distant relapse of disease is common which directly links with survival outcomes. The overall 5 year survival of TNBC is 81% [7]. However, with the development of metastases, the 5 year survival drops to 11% [7].

The challenge with TNBC is that it is a histological diagnosis of exclusion based on current immunophenotyping approaches. It is a heterogenous disease which requires multimodal stratification to optimise therapeutic selection. The diagnosis of TNBC could benefit from a two-step classification approach:

1. Classification of low-grade or high-grade TNBC
2. Molecular multimodal classification of high-grade TNBC

1.2.1 Low-Grade TNBC

Low-grade TNBC comprises 10% of TNBC cases [8]. There are three subtypes:

i. Adenoid cystic carcinoma (ACC)

ii. Secretory carcinoma

iii. Acinic cell carcinoma

ACC shares histological similarity with salivery gland ACC with low-grade proliferation and diffuse serous differentiation. ACC of the breast exhibits a high mutational burden and 80% have *TP53* mutations [8].

Secretory carcinoma typically presents in children and adolescents and rarely metastases outside of the breast. Histologically, secretions are present either within cysts similar to thyroid follicles or within luminal structures. Secretory carcinomas characteristically have a balanced t(12;15) which results in an *ETV6-NTRK3* gene fusion [8].

Acinic cell carcinoma is extremely rare. Its true incidence is unclear. It exhibits a characteristic clear 'hypernephroid' cytoplasm [8].

The natural history of low-grade TNBC is distinct to high-grade TNBC. It is clinically indolent [8]. Surgery is the mainstay of treatment with a limited role for chemotherapy [8]. Metastatic NTRK-fusion positive tumours are eligible for treatment with larotrectinib under the Cancer Drug Fund.

1.2.2 High-Grade TNBC

There are two commonly used molecular classification systems adopted in breast cancer:

i. Intrinsic molecular subtypes based on transcriptomics [9].

ii. Integrative clusters (IntClust) subtypes based on combined genomics-transcriptomics assessment [10].

I will focus on the intrinsic classification system as related to TNBC in the fol-

lowing section. The integrative clusters classification will be explored in Section 1.3.

The landmark work by Perou et al was the first molecular stratification of breast cancer using transcriptomic microarray-based profiling [9]. Five intrinsic breast cancer subtypes were identified: luminal-A, luminal-B, HER2-enriched, basal-like and normal breast-like [9].

Immunophenotypic TNBC samples are not restricted to a single intrinsic subtype. 75% - 85% of TNBC are classified to the basal-like subgroup, 5% - 15% to the HER2-enriched subgroup and the remaining cases are distributed amongst luminal-A, luminal-B and normal breast-like groups as shown in Figure 1.3 [11]. Therefore, TNBC consists of multiple different intrinsic subtypes.

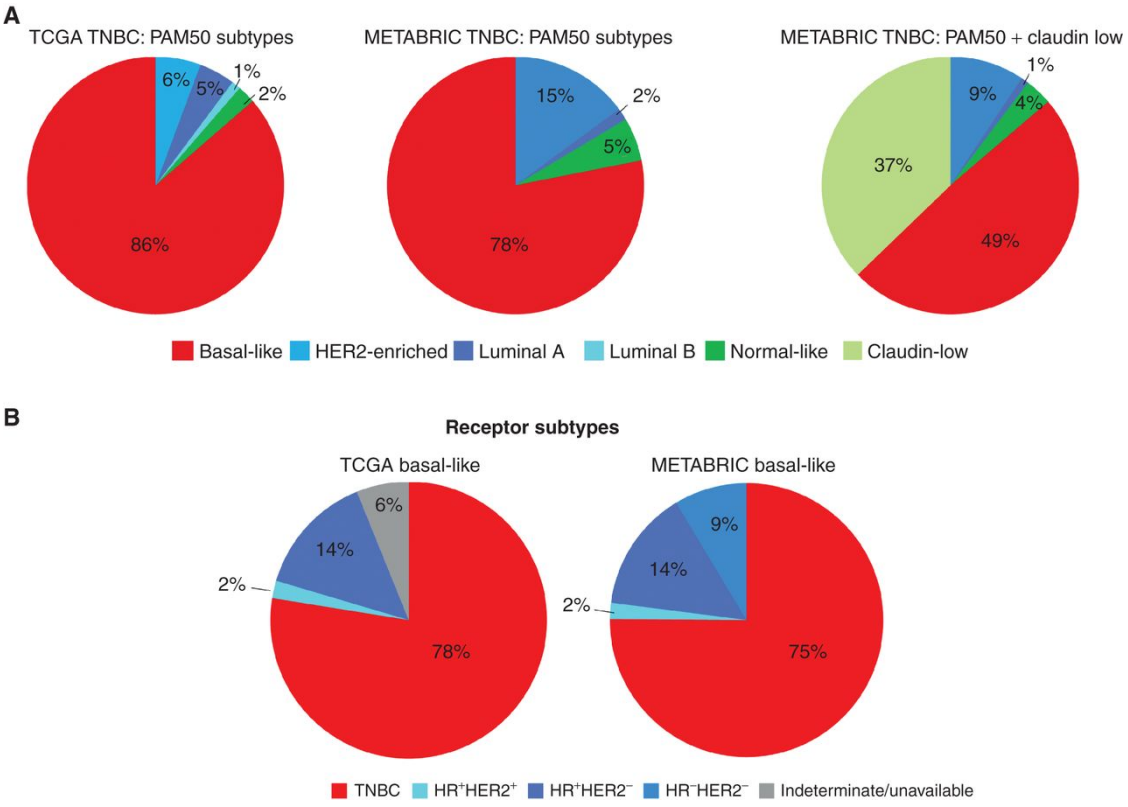


Figure 1.3: Relationship between TNBC and Intrinsic subtypes. a. Distribution of intrinsic breast cancer subtypes in TNBC from TCGA and METABRIC datasets. b. Distribution of receptor status in basal-like intrinsic subtype from TCGA and METABRIC datasets. Reproduced from [11].

Furthermore PAM50, a 50-gene assay predictive of complete pathological response (pCR) in most breast cancer subtypes, performs poorly when restricted to TNBC [12]. Application of PAM50 in TNBC is not predictive of pCR [12] indicative of the clinical challenges presented by the inherent heterogeneity of TNBC.

Therefore, neither the Intrinsic classification system nor currently available gene signatures offer adequate stratification for TNBC.

1.2.2.1 Molecular Stratification of TNBC

The first significant contribution to establishing a molecular taxonomy specific to TNBC was conducted by Lehmann et al [13]. Gene expression profiles from 587 human TNBC samples obtained across 21 published datasets were split into a training and validation set and investigated by K-means clustering.

Seven TNBC subtypes were identified by cluster-based differential gene expression analysis, namely:

1. Basal-like 1 (BL1)
2. Basal-like 2 (BL2)
3. Immunomodulatory (IM)
4. Mesenchymal (M)
5. Mesenchymal stem-like (MSL)
6. Luminal androgen receptor (LAR)
7. Unstable (UNS)

The molecular characteristics of the TNBC subtypes are summarised in Table 1.1.

Subtype	Enriched Gene Ontology	DE Genes	Drug Sensitivity
BL1	Cell cycle, DNA damage response, Proliferation	*AURKA*, *CCNA2*, *MYC*, *NRAS*, *CHEK1*, *FANCA*, *RAD51*, *MSH2*, *MKI67*	Cisplatin
BL2	Growth factor signaling, Glycolysis, Gluconeogenesis	*EGFR*,*MET*, *TP53*, *MME*	Cisplatin
IM	TH1/TH2 pathway, NK cell pathway, B cell and T cell receptor signaling, Antigen presentation	*JAK1*, *JAK2*, *LCK*, *LYN*, *NFKB1*, *STAT1*, *STAT4*, *ZAP70*	Nil
M	TGFβ signalling	*TGFB111*, *SMAD6*, *NOTCH1*, *TGFB1*, *TGFBR1*	Dasatinib
MSL	TGFβ signalling, EMT, Angiognesis, Low levels of proliferation markers and claudin	As per M subtype, *SNAI2*, *TCF4*, *TWIST1*, *KDR*, *TIE1*	Dasatinib
LAR	Steroid synthesis, Androgen/Oestrogen metabolism	*AR*, *DHCR24*, *APOD*	Bicalutamide, 17-DMAG
UNS	Nil	Nil	Nil

Table 1.1: Summary of TNBC Molecular Subtypes [13]

Intrinsic breast cancer subtypes were linked with representative TNBC cell lines in order to then explore subtype drug sensitivity. Potential therapeutics stratified according to subtype are shown in Table 1.1.

To date, molecular stratification of TNBC does not drive treatment selection in routine clinical practice, with the exception of patients with inherited predispositions to genomic instability. Following the results of the TNT trial, patients with germline BRCA1/BRCA2 mutations are treated with platinium-based chemotherapy rather than standard anthracycline-taxane based treatment [14].

There are limitations with the subtypes proposed in the Intrinsic classification system. Follow-up studies have not demonstrated consistent reproducibility of the subtypes, in particular BL2 [15]. The original study identified TNBC samples using gene expression. It has not been possible to reproduce the seven subtypes using IHC-defined TNBC samples [13]. The latter is a significant limitation since all routine clinical workflows utilise IHC for determination of receptor status.

More recent efforts to identify stable molecular TNBC subtypes have explored single cell techniques for TNBC characterisation. The potential of scRNA-seq to better deconvolve the interaction between the tumour and its microenvironment may enable identification of the dominant cell state driving tumour progression which should closely map with the underlying TNBC subtype, improve reproducibility with orthogonal assays and correlate better with therapeutic response and survival outcomes.

The work of Karaayvaz et al has demonstrated it is possible to recover previously established molecular TNBC subtypes at single cell resolution [16]. 1500 cells were recovered from six patients with localised, treatment naive TNBC [16]. For most tumours, multiple TNBC subtypes were identified per patient [16]. The subtype proportion varied across patients as shown in Figure 1.4.

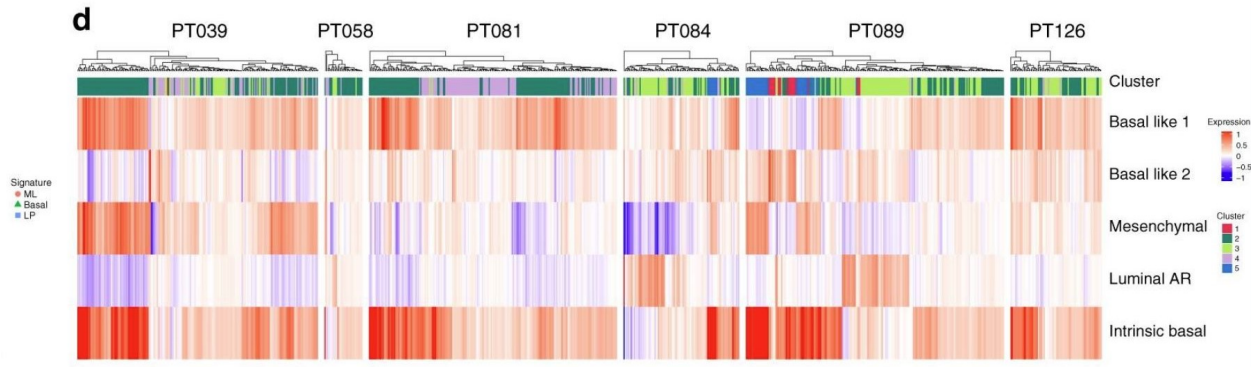


Figure 1.4: TNBC samples exhibit a combination of molecular subtypes. Reproduced from [16].

The findings provide further evidence for the extent of intratumoural heterogeneity unique to TNBC. It suggests bulk signatures may be poorly predictive since they are unable to capture such heterogeneity.

Karaayvaz et al offer a glimpse into the potential clinically predictive capacity of single cell signatures. A t-SNE plot of the malignant epithelial subpopulation is shown in Figure 1.5. Three previously established signatures of metastatic potential and treatment resistance [17, 18, 19] were found to be all independently enriched in a single cluster, cluster 2 as shown in dark-green, within the malignant epithelial subpopulation [16].

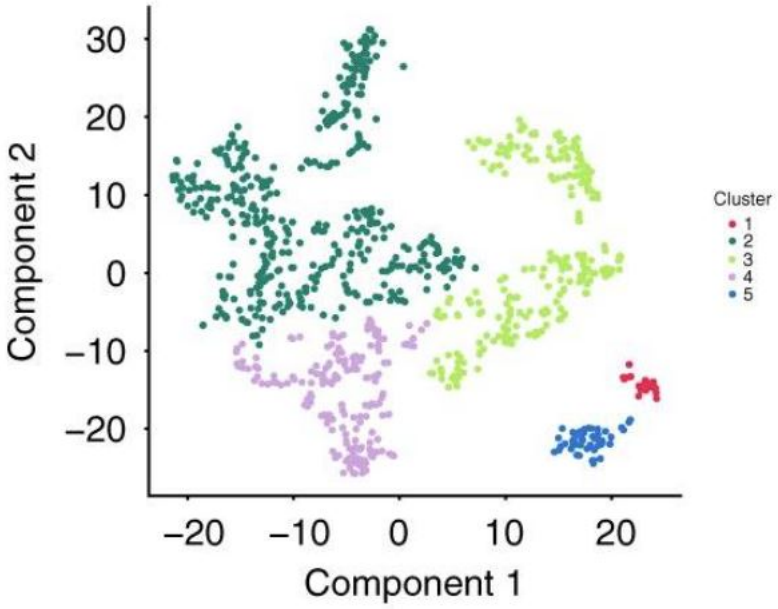


Figure 1.5: The malignant epithelial subpopulation exhibits transcriptional heterogeneity. Reproduced from [16].

The cluster-2 signature was found to be predictive for poor survival outcomes in the METABRIC dataset as shown in Figure 1.6. Pathway analysis revealed enrichment of programs relating to the glycosphingolipid pathway and innate immunity in the cluster-2 signature [16].

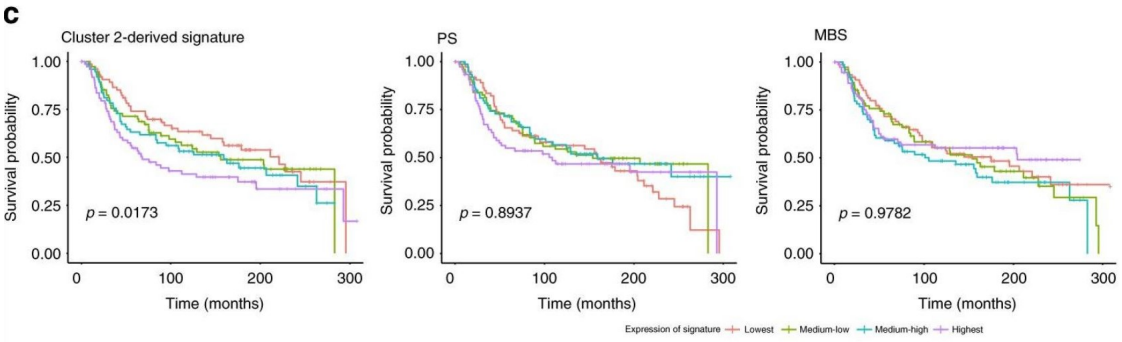


Figure 1.6: KM survival curves for TNBC patients from the METABRIC cohort. The cluster-2 signature is compared to a 70-gene prognostic signature (PS) and a 49-gene metastatic burden signature (MBS). Reproduced from [16].

The findings provide an initial framework to guide future work to refine the molecular stratification of TNBC, uncover further clinically relevant cell states

and importantly, identify consistent and reproducible multimodal programs which can guide clinical decision making.

1.3 The Molecular and Spatial Landscape of Breast Cancer

I will present an overview of the Integrative Clusters system, the latest molecular stratification framework for breast cancer followed by a discussion of the latest insights into the spatial organisation of breast cancer.

1.3.1 Molecular Landscape

The Intrinsic classification of breast cancer was a major advance [9]. However, it was limited to stratifying breast cancer based solely according to transcriptomics, its performance on non-transcriptomic based assays was limited and subtype reproducibility was not consistent.

Breast cancer is prototypically a disease dominated by copy number aberrations (CNA) [20]. Therefore, the combined integration of genomics, transcriptomic and clinical profiling from 1000 breast cancers (with an additional 1000 sample validation set) to identify clinically robust subtypes led to the development of the Integrative Clusters classification system [10].

A total of 10 subtypes, IntClust1 - IntClust10, were identified [10]. It is a significant contribution to the field, both in terms of the scale of data collected and in terms of its clinical validity. The classification system is robust [10]. The dual-modality classifier (ie combined genomic-transcriptomic classifier) was concordant with the unimodal classifier. Concordance was consistent when assessed on external test datasets [10].

The IntClust subtypes show a clear association with survival outcomes (Fig-

ure 1.7a) and with response to chemotherapy (Figure 1.7b). It outperforms stratification according to receptor status or intrinsic classification.

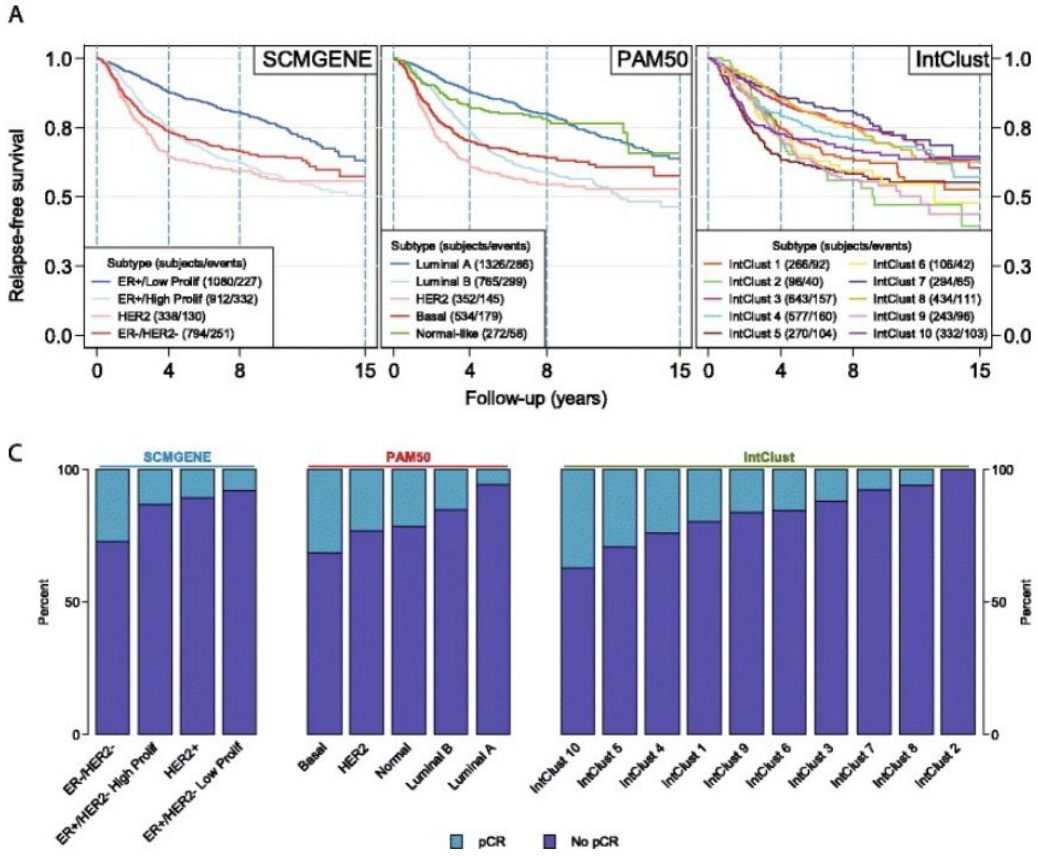


Figure 1.7: IntClust classification correlates with (a) survival outcomes and (b) chemotherapy sensitivity. Reproduced from [10].

IntClust also captures genomic drivers of breast cancer. Using breast cancer genes associated with CNA, it was found that variation in expression of these genes was best stratified according to IntClust subtypes. Therefore, IntClust captures important, clinically relevant aspects of breast cancer biology.

1.3.2 Spatial Landscape

The high-dimensional spatial profiling of breast cancer is rapidly evolving with the advent of new spatial-omics technologies which can be applied to clinical grade tumour samples. Several significant contributions have been made in the last two years by Jackson et al [21], Ali et al [22] and Danenberg et al [23].

It is a sequential pattern of advance in which each study builds upon and adds to the insights offered by its predecessors. The most recent study by Danenberg et al represents a culmination of our current understanding in the spatially-driven hierarchy in breast cancer. I will therefore focus discussion on the results presented by Danenberg et al, whilst recognising the earlier valuable contributions by Jackson et al and Ali et al.

The work presented by Danenberg et al is beautifully elegant. Samples from 700 patients previously enrolled to the METABRIC study were profiled on a 37-plex imaging mass cytometry (IMC) panel and combined with previously collected genomic and clinical data [23]. All patients were treatment naive and breast cancer subtypes, as defined across several different classification systems, were well-represented across the cohort.

8 distinct steps were explored during the course of investigation:

1. Identification of distinct epithelial and TME phenotypes
2. Assessment of the functional complexity of spatial architecture
3. Exploration of drivers of diversity
4. Detailed exploration of intercompartmental spatial boundaries

5. Exploration of the perivascular compartment
6. Identification of common multicellular structures using graph-based analysis
7. Exploration of the inderdependency between genomic and TME drivers
8. Correlation of survival outcomes with TME structures

1.3.2.1 Epithelial and TME Phenotypes

A total of 32 cell phenotypes, 16 epithelial and 16 TME, were identified. Phenotype identification was conducted as a two-step process [23]. A pretrained cytokeratin-based segmentation mask was first used to classify individually segemented cells as 'epithelial' or 'non-epithelial'. Clustering was then performed within each class type based on the IMC-captured protein expression data [23].

A summary of all phenotypes is shown in Figure 1.8. The heatmap on the left represents the epithelial phenotypes and the heatmap on the right represents the TME phenotypes. Phenotype labels are listed on the y-axis.

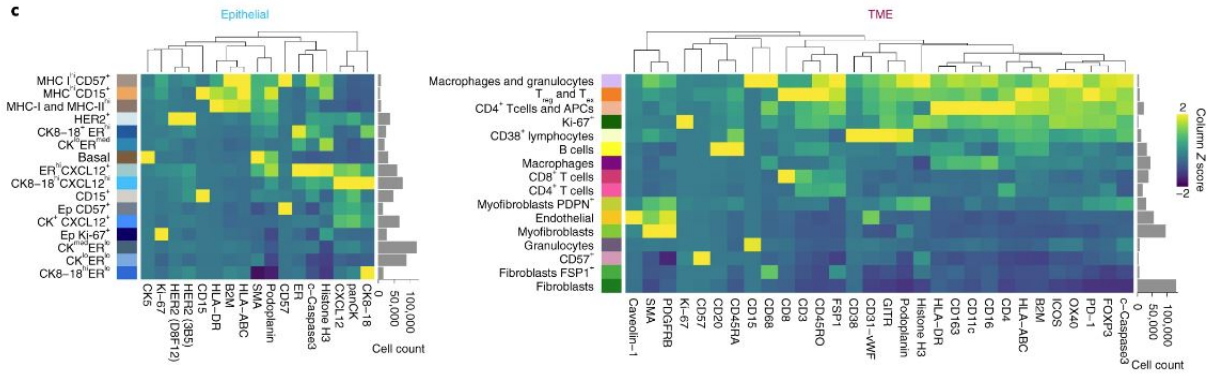


Figure 1.8: Epithelial and TME cell phenotypes. Reproduced from [23].

Epithelial phenotypes were generally defined based on cytokeratin and hormone receptor expression [23]. Three epithelial phenotypes were further defined by expression of antigen-presentation proteins (MHC-I +/- MHC-II) [23].

All major TME cell types consisting of lymphoid cells, myeloid cells and stromal cells, were well represented across the TME phentotypes [23]. Lymphoid cells were further subclassified into T_{helper}, $T_{cytotoxic}$ and T_{reg}, B cells

and $CD38^+$ cells. Myeloid cells were divided into macrophages, granuloctyes and dendritic cells. Fibroblasts were divided into myofibroblasts, $FSP1^+$ and $PDPN^+$ fibroblasts.

1.3.2.2 Functional Spatial Complexity

A metric based on Shannon's entropy was adopted to quantify phenotypic diversity between the epithelial and TME phenotypes across breast cancer subtypes [23].

The concept of entropy is rooted within Information Theory. Mathematically, information is defined as the amount of surprise associated with an outcome variable. For example, an unbiased coin showing heads has more surprise, ie it has more information, than a biased coin showing heads. Shannon's entropy is the average information associated with a probability distribution [24].

Entropy was generally greater for the TME phenotypes compared to the epithelial phenotypes as shown in Figure 1.9. TME entropy was highest for the IntClust4+ subtype. A notable exception is that epithelial entropy was high for the intrinsic basal subtype.

The observation of consistently higher entropy in the TME phenotypes suggests that entropic regulation of the tumour ecosystem is driven largely by the TME, a significant finding for this current era of TME-focused therapies such as immune checkpoint blockade.

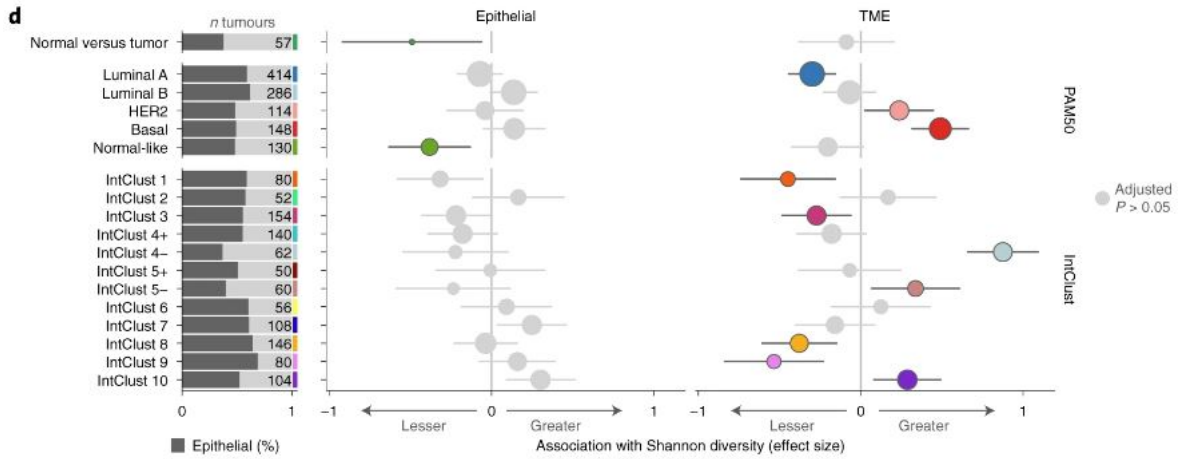


Figure 1.9: Comparison of Shannon entropy between epithelial and TME phenotypes across breast cancer subtypes. Reproduced from [23].

1.3.2.3 The Tumour-Stroma Interface

Investigation of the compositional complexity was explored in further detail by focusing at the spatial boundary between cell compartments [23]. Myofibroblasts were found to be enriched at the tumour-stroma interface with reciprocal depletion of lymphoid phenotypes, in particular B cells (Figure 1.10). These findings support the current model of myofibroblast-mediated immune exclusion of the epithelial tumour compartment [23].

1.3.2.4 The Vascular-Immune Interface

Detailed investigation was conducted on perivascular cells since the vascular network provides the transport framework for delivery of lymphoid cells into the tumour ecosystem. Perivascular cells were found to be most strongly enriched for endothelial cells along with $CD38^{+}$ lymphocytes and myofibroblasts (Figure 1.11). This suggests a model of lymphocytic adhesion to endothelial cells along with perivascular stromal activation [23].

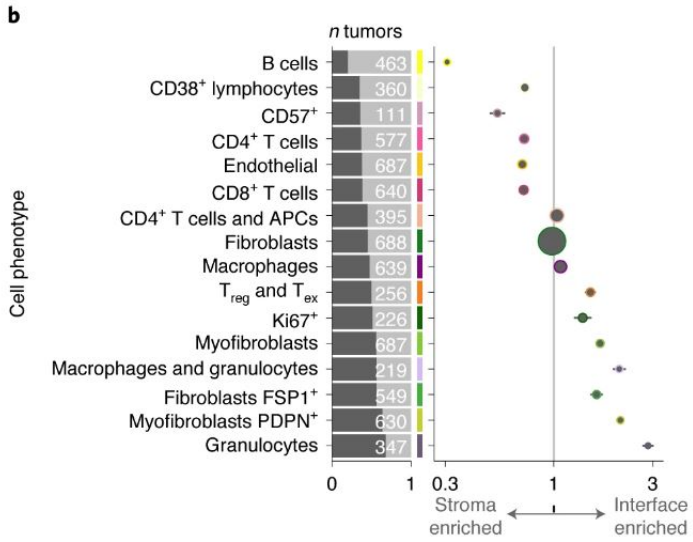


Figure 1.10: Cell types enriched at the tumour-stroma interface. Reproduced from [23].

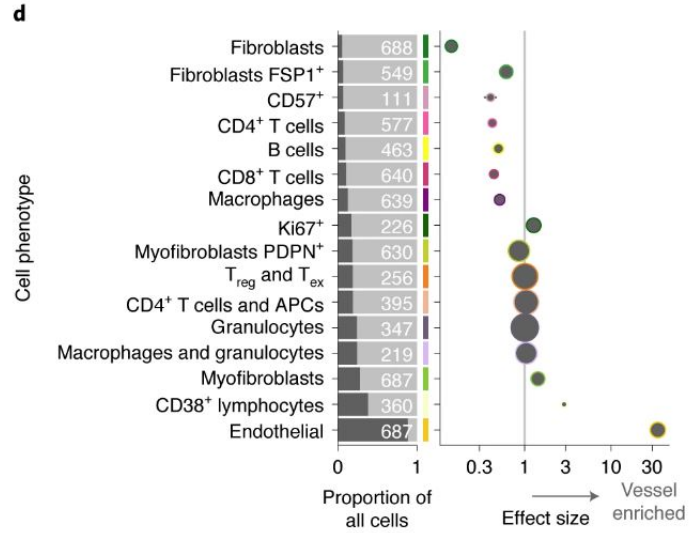


Figure 1.11: Cell types enriched at the perivascular interface. Reproduced from [23].

1.3.2.5 Multicellular TME Structures

By applying community detection algorithms to graph-based representations of the IMC data, 10 multicellular spatial TME structures preserved across samples were identified as shown in Figure 1.12.

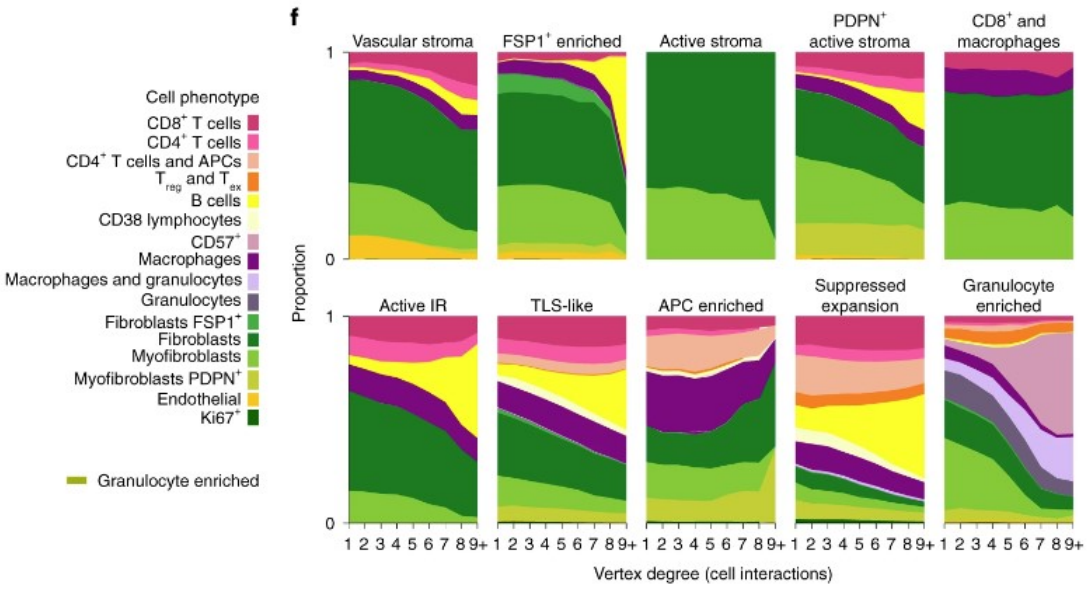


Figure 1.12: Multicellular TME structures. Reproduced from [23].

Table 1.2 shows the list of TME structures. The structures differ based mainly on stromal and leukocyte composition [23].

Vascular stroma
FSPR$^+$ enriched
Active stroma
PDPN$^+$ active stroma
5CD8$^+$ and macrophages
Active IR
TLS-like
APC enriched
Suppressed expansion
Granulocyte enriched

Table 1.2: Spatially Preserved TME Structures [23]

All TME structures contain fibroblasts and myofibroblasts to varying degrees. Active IR, TLS-like and suppressed expansion classes are notably enriched for B cells. Granulocyte enriched is the only class containing granulocytes.

The presence of T_{reg} cells in the suppressed expansion and granulocyte structures suggest these structures may represent dysfunctional immune states. It would be expected patients enriched for these structures may be less likely to benefit from immune checkpoint based therapies.

I think the most interesting observation is the consistent presence of fibroblasts across all structures. Although not a causal link, it suggests the fibroblast compartment is integral to shaping the TME. Detailed mechanistic investigation into the tridirectional relationship between fibroblasts, lymphoid and epithelial cells may uncover new mechanisms to target tumour vulnerabilities. Fibroblasts should not be a forgotten compartment.

To date, no fibroblast-targeting therapies are used in clinical practice. Several promising candidates, such as RO6874281 and ABBV-085, are in early phase clinical trial development (NCT02627274 and NCT02565758 respectively).

The dynamics between genomic drivers of breast cancer and TME structures are complex. Broadly, breast cancer subtypes exhibited distinct patterns of enrichement of TME structures. Enrichment patterns were most strongly associated with ER status.

The suppressed expansion structure, an immunosuppressive structure, was enriched in ER^- tumours regardless of HER2 expression [23]. BRCA1 and CASP8 mutations were associated with the suppressed expansion structure [23].

It is currently unclear if these mutations reinforce or are a compensatory response to the immunosuppressive consequences of the suppressed expansion structure. BRCA1 mutations result in a deficiency in homologous recombination mediated DNA repair. The resulting mutational signature may stimulate an adaptive immune response [25]. In contrast, CASP8 mutations protect against CD8 T cell mediated apoptosis [26].

Mutations in *CDH1* are associated with $PDPN^+$ active stroma and vascular stroma [23]. Mutations in *CDH1* are commonly present in invasive lobular breast cancer [23]. Lobular breast cancer typically exhibits a diffusely infiltrative pattern of single file growth. Therefore, cellular dynamics at the tumour-stroma interface may account for the distinct spatial organisation of lobular breast cancer.

1.3.2.6 Spatially Resolved TME Structures correlate with Clinical Outcomes

Four TME structures were associated with survival outcomes in ER+ disease [23]. Vascular stroma was associated with better survival outcomes [23]. Granulocyte enriched, APC enriched and suppressed expansion were poor prognostic TME structures [23]. Enrichment of T_{reg}, as present in the suppressed expansion structure, has been previously linked to worse survival outcomes [27].

No TME structures were prognostic for ER– disease [23].

1.3.2.7 Limitations and Conclusion

The study was powered to offer insight into correlations between specific patterns of TME structural enrichment and overall clinical outcomes. Unravelling the underlying causal relationships will be challenging due to the presence of multiple confounding factors in what is a temporally changing biological system subject to a wide range of internal and external perturbations.

Overall, the insights offered by Danenberg et al are an important and valuable contribution to the field. It demonstrates the potential advances which can be offered by spatial-omics technologies. I anticipate the future intention will be to integrate such high-dimensional spatial profiling with already established multimodal ensemble-based regressors in breast cancer [28] in order to tailor and refine their predictive capacity. It represents the early stages of the next generation of medicine.

1.4 Intention of the DPhil Project

My original intention for my DPhil was similar to several components of the very recently published work by Danenberg et al.

I intended to conduct Slide-seq based spatial transcriptomic profiling, scRNA-

seq and LCM-based spatial proteomics in a cohort of 30 treatment naive TNBC clinical samples. As part of the study, I also intended to directly compare two new spatial transcriptomic technologies, Slide-seq and CARTANA in-situ sequencing and contribute to the development of biologically-interpretable computational methods for spatial-omics datasets.

2

Experimental Design

Contents

2.1 Technologies Available at the Outset of the DPhil 26
2.2 Intended Technology Deployment 26
2.3 Critical Analysis of Spatial Transcriptomic Techniques 27
2.3.1 Overview of Techniques 27
2.3.2 Selection of Techniques 30
2.4 Protocols 31
2.5 Pandemic-related DPhil Disruption 36

2.1 Technologies Available at the Outset of the DPhil

The DPhil explored the application of two techniques:

1. Spatial transcriptomic technologies
2. Dissociated single cell sequencing technologies (scRNA-seq)

Both techniques were not available or established at the beginning of the DPhil. There was no pre-existing expertise in the development and application of these techniques. The development of the experimental and bioinformatic pipelines was to form a core part of the DPhil. In such a context, the aim was to develop the pipelines in collaboration with scientists experienced in their application.

I established de novo all academic and commercial collaborations, scientific and financial agreements and IP rights as part of the DPhil.

2.2 Intended Technology Deployment

The original DPhil intention was to conduct multimodal profiling using spatial transcriptomics and scRNA-seq in a collection of thirty breast cancer samples donated from patients in the treatment naive setting. Transcriptomic profiling was to be paired with proteomic profiling using LCM-based proteomics and traditional immunohistochemical approaches.

scRNA-seq was to be completed using the 10X Genomics Chromium platform and spatial transcriptomcs was to be completed using Slide-seq. scRNA-seq was to be performed in collaboration with Dr. Thomas Carroll based in the lab of Prof. Xin Liu, Ludwig Institute, Oxford. Slide-seq was to be performed in collaboration with Dr. Evan Murray, a research scientist at the Broad Institute, Boston.

A comparison between two conceptually different spatial transcriptomic techniques, Slide-seq and CARTANA in situ sequencing, was to be conducted

in a subset of the patient samples. CARTANA in situ sequencing was provided through the CARTANA in-house technical team based in Sweden.

An established pipeline for LCM-proteomics was available in collaboration with Dr. Roman Fischer based at the TDI, University of Oxford. I developed expertise in performing laser microdissection to collect regions of tissue which could then later be processed in the Fischer lab.

2.3 Critical Analysis of Spatial Transcriptomic Techniques

2.3.1 Overview of Techniques

Spatial transcriptomic techniques can be divided into three general approaches:

i. Laser microdissection of regions of interest

ii. Imaging-based ST

iii. Sequencing-based ST

2.3.1.1 Laser Microdissection

Microdissection can be performed using laser capture (LCM) [29] [30]. Tissue sections are prepared on uncharged membrane glass slides which facilitates tissue extraction of dissected regions [29] [30]. Tissue is stained with cresyl violet to aid visualisation during dissection [30]. The laser microdissection unit incorporates a microscopy unit so that regions of interest can be identified and catapulted onto PCR tubes for RNA capture [29] [30].

This brute-force approach offers low sample-throughput and is not amenable for wide scale coverage of tissue. High resolution capture is particularly problematic since it is difficult to achieve reliable extraction of single cells. The

downstream RNA extraction process for low input RNA also requires protocol optimisation in a tissue specific manner [29]. There is therefore a need to develop high-throughput multiplexed approaches using a range of imaging- and sequencing-based ST.

2.3.1.2 Imaging-Based ST

Imaging-based ST enable direct visualisation of RNA molecules within their native environment [31]. Tissue is fixed after which mRNA is first reverse transcribed to cDNA [31]. Fluorescently labelled padlock probes complementary to the target of interest can then bind to the generated cDNA [31]. Target amplification, required for downstream detection, is achieved by a process called rolling-circle amplification (RCA) [31].

RCA generates multiple copies of the padlock probe with inter-probe gaps filled by sequencing-by-ligation [31]. The end product is imaged and decoded in a manner analogous to Illumina sequencing-by-synthesis. CARTANA provides commercial solutions to access multiplexed in situ hybridisation techniques [32].

2.3.1.3 Sequencing-based ST

Sequencing-based ST involves in situ transcript capture followed by ex situ sequencing. Tissue is fixed, stained, imaged and permeabilised onto the array [33]. Permeabilised mRNA hybridises to the barcoded primers and in situ RT is performed [33]. The barcoded reads are then mapped back to the original tissue location following completion of library preparation and next generation sequencing [33].

There are several sequencing-based ST products: Visium, Slide-seq, High-Definition Spatial Transcriptomics (HDST) and the GeoMx Digital Spatial Profiler (DSP).

The 10X Visium solution consists of a glass slide onto which a barcoded RT primer is printed [33]. The barcoded primer encodes the x and y coordinate position on the array [33]. Visium offers limited spatial resolution. Spatially profiled regions were 55 μm diameter circular spots as per the technology available in 2021.

Slide-seq is a high resolution spatial transcriptomic technology first published in 2019 for its application in murine brain tissue [34]. Polystyrene beads, each of 10 μm diameter, are attached in monolayer onto a glass coverslip in place of printing barcoded RT primers onto a glass slide [34]. Each bead has a randomly attached unique spatial barcode [34]. The bead location is not known a priori. The bead position is decoded in situ using sequencing-by-ligation [34].

Slide-seq spatial resolution is governed by the bead diameter. The spatial resolution is greater than Visium. At 10 μm resolution, Slide-seq enables the spatial profiling of a small number of single cells.

HDST is very similar technique to Slide-seq. In HDST, 2 μm diameter spatially profiled beads are deposited onto a glass slide [35]. Slide-seq and HDST are only suitable for use on fresh frozen tissue [34] [35].

An analogous approach suitable for FFPE is provided by the Nanostring GeoMx DSP [36]. RNA probes linked with a photocleavable barcoded tag are added to the tissue section [36]. User defined regions of interest consisting of 600 μm diameter circles are excited with UV light, enabling release of the RNA probes [36]. The released read is quantified using the NanoString nCounter instrument [36].

In summer 2019, a targeted GeoMx DSP 200 tumour-immune gene panel was available. The GeoMx Spatial Profiling instrument was due to be available in Oxford by early 2020.

2.3.2 Selection of Techniques

The available spatial transcriptomic platforms were explored on the basis of the following factors:

i. Suitability for use on breast cancer tissue

ii. Application in FFPE or fresh frozen tissue

iii. Availability of equipment

iv. Cost

A comparison of the benefits and disadvantages of the available spatial transcriptomic technologies is summarised in table 2.1

Technology	Tissue Type	Availability	High-throughput	Cost-effective
Slide-seq	Fresh frozen	✓	✓	✓
CARTANA	Fresh Frozen, FFPE	X	✓	X
LCM	Fresh Frozen, FFPE	✓	X	✓
GeoMx DSP	Fresh Frozen, FFPE	X	X	X

Table 2.1: Comparison of Spatial Transcriptomic Technologies

Following a literature review and discussion with scientists with hands-on experience with spatial transcriptomic technologies, two spatial transcriptomic platforms were selected for application in breast cancer during the DPhil:

1. Slide-seq
2. CARTANA in-situ sequencing

2.3.2.1 Slide-seq

Slide-seq offers high-resolution spatial transcriptomic profiling of fresh frozen tissue without requiring expensive, bespoke equipment or microscopes. It is suitable for high-throughput work and it is a cost-effective alternative to Visium.

2.3.2.2 CARTANA in-situ sequencing

Most spatial transcriptomic options to date are suitable for use only on fresh frozen tissue. The use of spatial transcriptomics in FFPE tissue presents additional limitations due to the RNA degradation which occurs during tissue processing. CARTANA was one of the few spatial transcriptomic options suitable for use on FFPE tissue in 2019.

2.4 Protocols

2.4.0.1 Slide-seq Protocol

Slide-seq pucks were produced in the Macosko lab in Boston, USA and then shipped to Oxford. 10 μm tissue sections were prepared and placed onto the circular 3 mm diameter spatial capture areas. The capture areas contain 10 μm polystyrene beads linked with a spatially-unique oligonucleotide.

Tissue permeabilisation and RNA hybridisation was completed for a range of time durations from 15 minutes to 105 minutes. Reverse transcription, sec-

ond strand synthesis, bead clean up (Ampure XP beads, Beckman Coulter, catalogue number A63881) and PCR amplification were performed with intervening washes.

Barcoded libraries were constructed using the Nextera XT kit (Illumina, catalogue number FC-131-1096) as per the manufacturer's instructions. A range of custom primers were used as detailed in Table 2.2:

Primer Name	Custom Sequence
Template Switch Oligo	AAGCAGTGGTATCAACGCAGAGTGAATrG+GrG
Truseq PCR Handle	CTACACGACGCTCTTCCGATCT
SMART PCR Primer	AAGCAGTGGTATCAACGCAGAGT
Truseq P5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTC
dN-SMRT	AAGCAGTGGTATCAACGCAGAGTGANNNGGNNNB

Table 2.2: Slide-seq Custom Primers

Libraries were quantified using the Qubit dsDNA High Sensitivity assay (Invitrogen, catalogue number Q32854) as per the manufacturer's instructions. Library QC was performed on the Agilent High Sensitivity dsDNA chip (Agilent, catalogue number 5067-4626) using the Agilent 2100 Bioanalyzer instrument located in the LICR Oxford branch.

Final libraries were diluted to a concentration of 4 nM and denatured for analysis using 75 base-pair pair-end reads. Samples were loaded on the NSQ 500/550 High Output kit v2.5 75 cycles (Illumina, catalogue number 20024906). Samples were sequenced using the Illumina NextSeq 500 Walk-In sequencing service at the Oxford Genomics Centre.

2.4.0.2 Slide-seq Data Processing

Raw sequencing files were transferred to the Macosko lab. The pre-processing steps were completed by the Macosko lab using a Slide-seq specific bioinformat-

ics pipeline internal to the Macosko lab. The analysis tools included Illumina barcode extraction, base-calling, alignment to a human reference genome, post-alignment QC and matching of Illumina barcodes to bead barcodes.

2.4.0.3 CARTANA Protocol

FFPE breast cancer tissue blocks were transferred to the CARTANA In Situ technical team, Sweden. Optimisation of the hypoxia probes and application of the novel hypoxia panel in FFPE breast cancer samples was completed by the CARTANA technical team using a proprietary internal protocol. The generated raw data was shared by the CARTANA technical team.

2.4.0.4 Droplet based scRNA-seq

Fresh tissue biopsies were enzymatically dissociated using an unpublished protocol. The protocol had been optimised for clinical tissue dissociation in preparation for single cell droplet encapsulation. It was kindly shared by the Single Cell Genomics team at the Cold Spring Harbor Laboratory, USA.

Upon dissociation, red blood cells were removed using 10ml of red blood cell lysis solution (Miltenyi, catalogue number 130-094-183). Dead cells were removed using the Miltenyi Dead Cell Removal kit (Miltenyi, catalogue number 130-090-101) as per the manufacturer's instructions. Cell viability was assessed on the Countess II Automated Cell Counter. Following viability check, cells were re-suspended at 1000 cells per μl.

Cells were then transferred on ice to the Single Cell Genomics facility at the WIMM. Cells were loaded onto the 10X Chromium platform at a concentration of 1000 cells per μl as per the manufacturer's instructions. 10,000 cells were loaded per experiment per patient sample.

Single cell encapsulation was completed on the 10X Chromium Next GEM

Single Cell gene expression 3' kit (10X Genomics, catalogue number 1000269) and 10X Chromium Next GEM Single Cell gene expression 5' kit (10X Genomics, catalogue number 1000425). Generation of single cells in gel beads in emulsion (GEMs), barcoding, reverse transcription clean up, cDNA amplication and library quantification were performed as per the manufactuer's instructions.

Libraries were quantified using the KAPA-Illumina PCR quantification kit (Roche, catalogue number KR0405). Library QC was performed on the Agilent High Sensitivity dsDNA chip (Agilent, catalogue number 5067-4626) on random wells and with negative controls using the Agilent 2100 Bioanalyzer.

Final libraries were diluted to a concentration of 4 nM and denatured for analysis using 150 base-pair pair-end reads. Samples were loaded on the NSQ 500/550 Hi Output kit v2.5 150 cycles (Illumina, catalogue number 20024907). Samples were sequenced using the Illumina NextSeq 500 Walk-In sequencing service at the Oxford Genomics Centre.

2.4.0.5 scRNA-seq Data Processing

QC on raw sequence reads was performed using FastQC software. The 10X Genomics Cell Ranger version 5.0.1 software was used to process, align and calculate unique molecular identifier (UMI) counts against the human hg38 reference genome.

Count matrices were imported into R v4.0.0 on an Ubuntu operating system v18.04 and v20.04. Seurat v3.1.3 was used to perform data preprocessing, identification of highly variable genes, principal component selection, neighborhood-graph based clustering and UMAP visualisation. Harmony v1.0 was used for merging post-processed samples between experimental batches [37].

2.4.0.6 Immunohistochemistry

4 mm FFPE tissue sections were prepared on the Leica rotary HistoCore microtome. 4 mm fresh frozen tissue sections were prepared on the Bright Instruments OTF-5000 cryostat. Tissue sections were placed on SuperFrost Plus adhesion glass slides (Thermo Fisher, catalogue number 10149870).

FFPE slides were firstly dewaxed in HistoChoice clearing agent (Sigma, catalogue number H2779) and then rehydrated using decreasing concentrations of ethanol (Sigma, catalogue number 1070174000) to tap water.

For IHC staining, slides were stained using the FLEX staining kit (Agilent Dako Envision kit, catalogue number K8023). Antigen retrieval was performed in pH 6 for CA9 and GLUT1 antibodies by autoclave heating (56 °C for 10 minutes). Antigen retrieval was performed in pH 9 by autoclave heating (56 °C for 10 minutes) for the pimondazole antibody.

Endogenous peroxidase activity was blocked. Slides were then stained with the following primary antibody dilutions for 1 hour at room temperature:

i) Pimonidazole: mouse, Hypoxyprobe-1, Chemicon International, 1:100.

ii) CA9: mouse, M75, absolute antibody, 1:100.

iii) GLUT1: rabbit, ab166618, Abcam; 1:100.

Slides were washed in Flex buffer (Thermo Scientific, catalogue number 8101) and then incubated with Flex anti-rabbit/anti-mouse secondary antibody (Abcam, catalogue number ab205719) for 30 minutes at room temperature. Slides were then washed in Flex buffer.

3,3-Diaminobenzidine (Flex-DAB) was applied to the sections for 10 minutes (Agilent, catalogue number C80611-2). The slides were counter-stained by immersing in Flex-hematoxylin solution for 5 min, washed and air-dried before being mounted with mounting medium (Sigma, catalogue number 06522). Secondary-only control staining was routinely done and they were negative.

Slides were scanned on the Hamamatsu NanoZoomer 2.0-HT slide scanner located in the Histopathology department, John Radcliffe hospital, Oxford. Slides were visualised with the NDP.view2 image viewing software (version U12388-01).

Histological work was conducted in the histology lab of the Nuffield Division of Clinical Laboratory Sciences (NDCLS), University of Oxford with the exception of cryosectioning.

Cryosectioning was conducted in the histology lab of the Ludwig Institute of Cancer (LICR) Research Oxford branch with the prior consent of Dr. Stan Ng, LICR Facilities Manager. Upon completion of cryosectioning, fresh frozen tissue sections were transported on dry ice to the NDCLS, John Radcliffe hospital and stored in a HTA-compliant −80 °C freezer. Clinical and xenograft samples were tracked. A transfer log was maintained during transport between research sites.

2.5 Pandemic-related DPhil Disruption

The DPhil plan experienced disruption as a consequence of the COVID-19 pandemic.

Access to critical equipment housed across different locations between different research institutions was not permitted due to covid-related working restrictions resulting in significant experimental delays. As a consequence of the disruption, key collaborations discontinued.

The final phase of the DPhil involved multiple attempts to re-establish spatial transcriptomics pipelines.

3

Application of Spatial-omics Technologies in Breast Cancer

Contents

3.1 Opportunities of Spatial Biology in Cancer Medicine 39
3.1.1 The Application of Spatial-omics 39
3.1.2 Insights offered by Spatial-omics 46
3.1.3 Opportunities in Data Interpretation 57
3.1.4 Computational Tools 58
3.2 Methods and Results 62
3.2.1 Experimental Pipelines for Spatial Sequencing 62
3.2.2 Slide-seq: Experimental Pipeline 63
3.2.3 Slide-seq: Optimisation for Solid Tumours 64
3.2.4 Slide-seq: Application to Human Tissue 68
3.2.5 Slide-seq: Results 72
3.2.6 CARTANA: Commercial Pipeline 75
3.2.7 CARTANA: Bespoke Spatial Hypoxia Panel 77

3.2.8 CARTANA: Results 78
3.3 Challenges and Limitations 82
3.3.1 Quality Control 82
3.3.2 New Computational Tools 87
3.3.3 Limitations 89

3.1 Opportunities of Spatial Biology in Cancer Medicine

Spatial-omics will be the next frontier in biology, offering new insights into the functionally relevant cross-talk relationships which govern normal and aberrant tissue development [38]. It enables gene expression detection whilst retaining the spatial location of RNA transcripts or protein[38].

The opportunities and challenges of spatial-omics can be subdivided into:

i. The application of spatial-omics

ii. The insights offered by spatial-omics

iii. Opportunities in spatial-omics data interpretation

iv. Tools essential to maximising the biological insight offered by spatial-omics

3.1.1 The Application of Spatial-omics

I foresee that the application of spatial-omics in tumour biology will encompass two distinct phases.

Phase 1 entails a formal characterisation of the spatial dependency in gene expression across a range of tumour types (squamous cell carcinoma vs adenocarcinoma) arising from different anatomical sites (eg colon, ovary, breast) at different stages in the treatment pathway (prior to treatment, during treatment, at local or distant relapse, in the heavily pre-treated setting).

To date, detailed investigation of spatial location as a biological and clinically relevant covariate has been limited by technological constraints. As a starting point, the ST field should first catalogue a core set of spatial structures preserved and reproduced across a range of ST technologies. In essence, a topological set preserved within transcriptomic space is being generated.

Phase 2 would build upon phase 1 to quantify intra- and inter-tumour spatial heterogeneity in 2D and 3D physical space using a greatly expanded sample

collection. I anticipate that the phase 2 workstream would best be conducted within a Human Cell Atlas (HCA)-linked initiative.

The results of phase 1 and phase 2 can be considered from a short-term and long-term perspective. In the short-term, prediction of gene expression from histopathology images (and vice versa) would be a natural objective. Exploration of spatial location as a causal factor in driving and shaping the subclonal evolution of cancer could also be pursued [39].

From a long-term perspective, the goal would be to identify all short and long-range spatial dependencies preserved across tumour types and understand how these dependencies determine response to internal microenvironmental perturbations (e.g. changes to intra-tumoural blood flow) and external perburbations (e.g. therapy). Spatially dependent structures could then be identified early during treatment to predict for response to different therapeutic classes as demonstrated through the work of Hwang et al [40].

The potential discovery opportunities and clinical benefits which could be unlocked by spatial-omics is shown by the cutting-edge work of Lomakin et al [41].

Multiregion sampling from two patients with multifocal ER positive and triple negative breast cancer was performed [41]. Both patients had background DCIS. A new workflow, BaSISS was developed on serial fresh frozen sections to enable spatial clonal mapping and spatial phenotyping using in-situ sequencing, bulk WGS and traditional IHC as shown in Figure 3.1 [41].

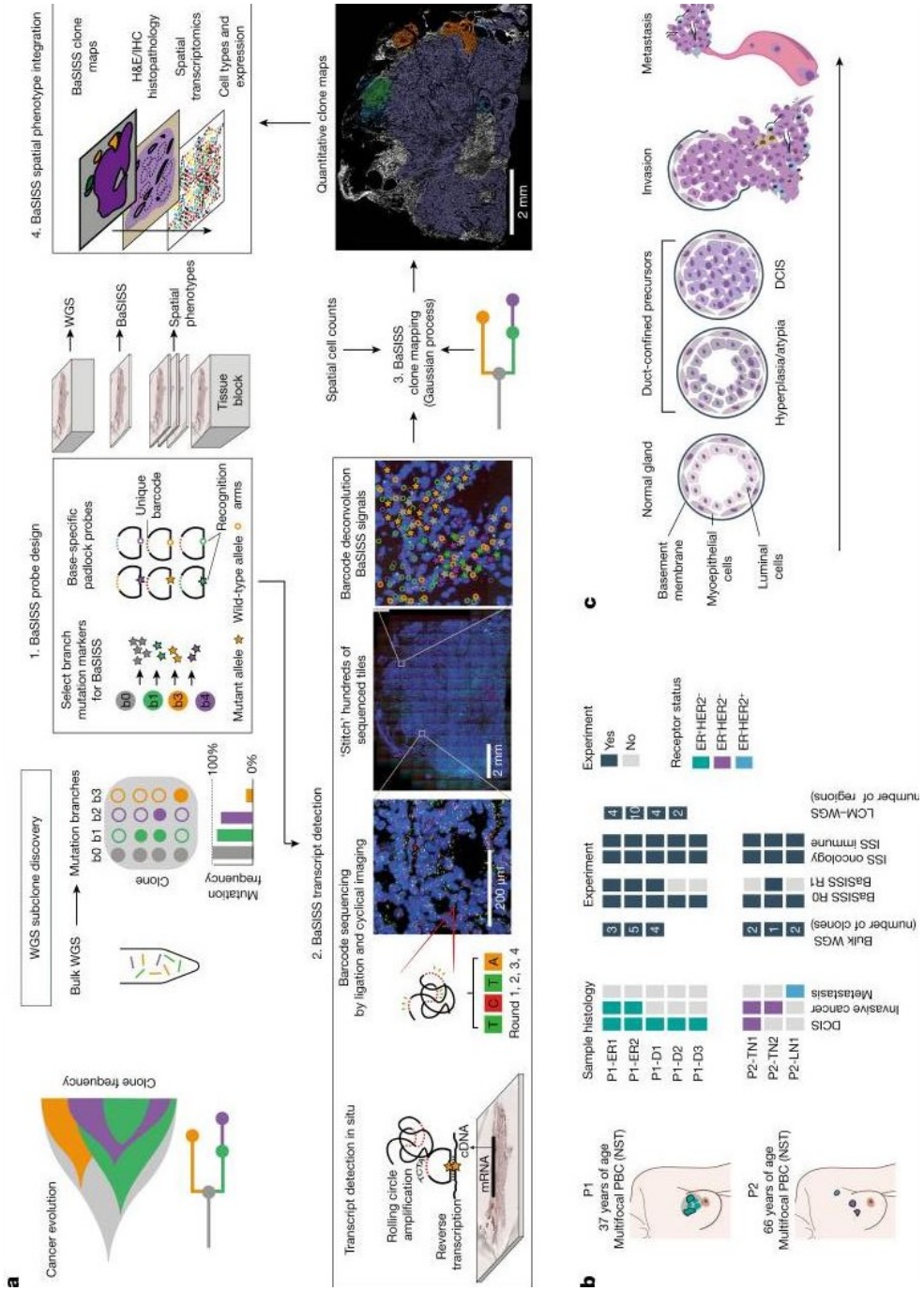


Figure 3.1: BaSISS Workflow. Reproduced from [41].

Spatial clone maps were developed for regions of invasive or in-situ disease (Figure 3.2) [41].

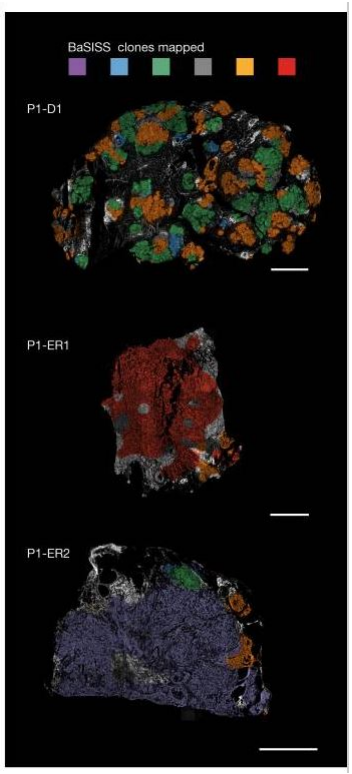
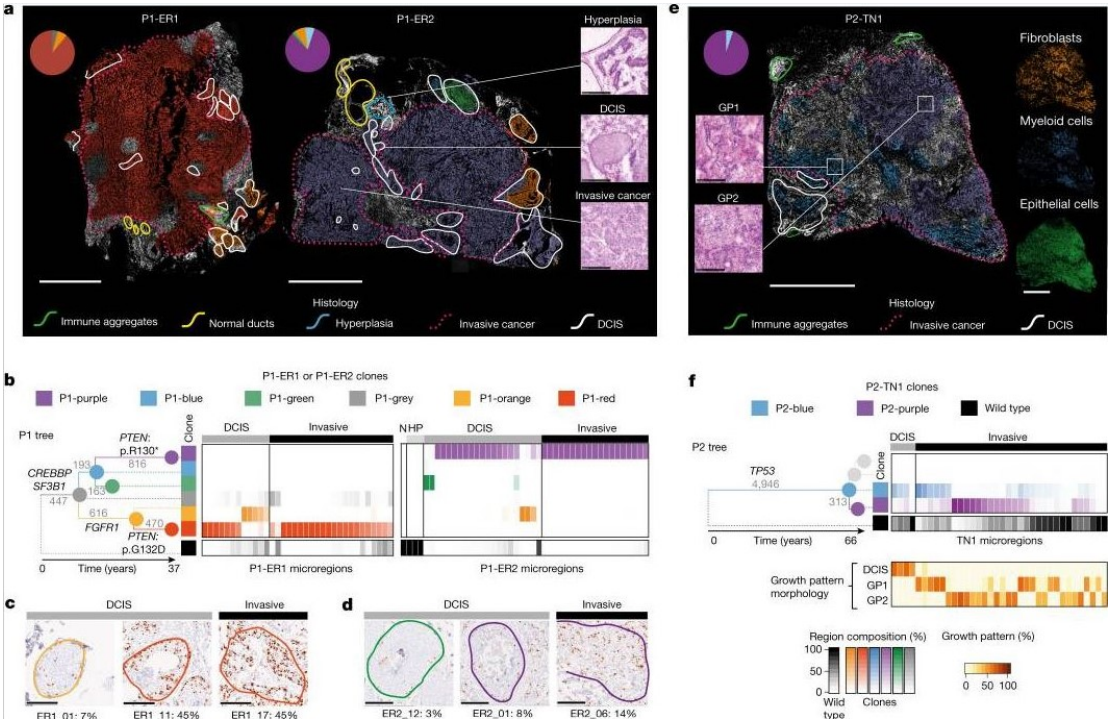


Figure 3.2: Spatial Clone Maps in Breast Cancer. Reproduced from [41].

The relationship between the spatial pattern of subclones and histologic structure was found to be broadly consistent [41]. Most sites of invasive disease were found to arise from the most recently diverged subclone [41].

Interestingly, regions of divergence were identified. In each sample, a subclone was identified which spanned both DCIS and invasive disease [41]. It suggests that the mutations required for invasive transformation originate within the ducts and thereby allow for stromal invasion (Figure 3.3ab and 3.3af) [41].

Divergence between biological modalities was investigated in further detail for PTEN mutant clonal regions in ER positive disease. PTEN mutant clones exhibited higher Ki-67 staining compared to PTEN wildtype region (Figure 3.3ac and 3.3ad) [41]. PTEN mutant clones in DCIS were compared with PTEN mutant clones in regions of invasive disease using ST [41]. Epithelial cell genes were differentially expressed between DCIS and invasive disease, specifically ACTB, KRT5 and CTSL2 and may provide insight into the causative factors driving histological transition [41].



(a) Spatial Relationship between Genetic and Histologic Structure. Reproduced from [41].

An important component of the study was investigation into the spatial patterns of growth for nodal metastatic disease [41]. Nodal metastases are a predictive marker for development of distant relapse, the major cause of death for cancer patients across disease sites.

Two spatially distant clones were identified in the involved node of a patient with TNBC: P2-blue and P2-orange (Figure 3.4a) [41]. Only the clone P2-blue was identified at the primary site of disease [41].

Each clone was associated with distant patterns of histologic growth. Clone P2-blue clustered around dense lymphocytic cores (Figure 3.4c) [41]. In contrast, clone P2-orange formed pseudo-linear sheets alongside sinusoidal structures (Figure 3.4c) [41].

These structural differences were mirrored with associated transcriptional differences, revealing that distinct subclones are present within distinct microenvironments (Figure 3.4e - 3.4g) [41].

P2-blue clones cluster around B-cell rich germinal-like centres, showing a potential clone-specific adaptive immune response [41]. P2-orange clones were found within macrophage-rich lymphoid sinuses [41]. CXCL8 was the most highly enriched gene in this region [41]. CXCL8 is preferentially released by hypoxic macrophage, thereby suggesting adaption of the P2-orange clone to hypoxic coditions [41].

These results show how clones can be spatially regulated by micrenvironmental conditions and supports the need for expanding the available portfolio of tumour microenvironment-targeting therapies.

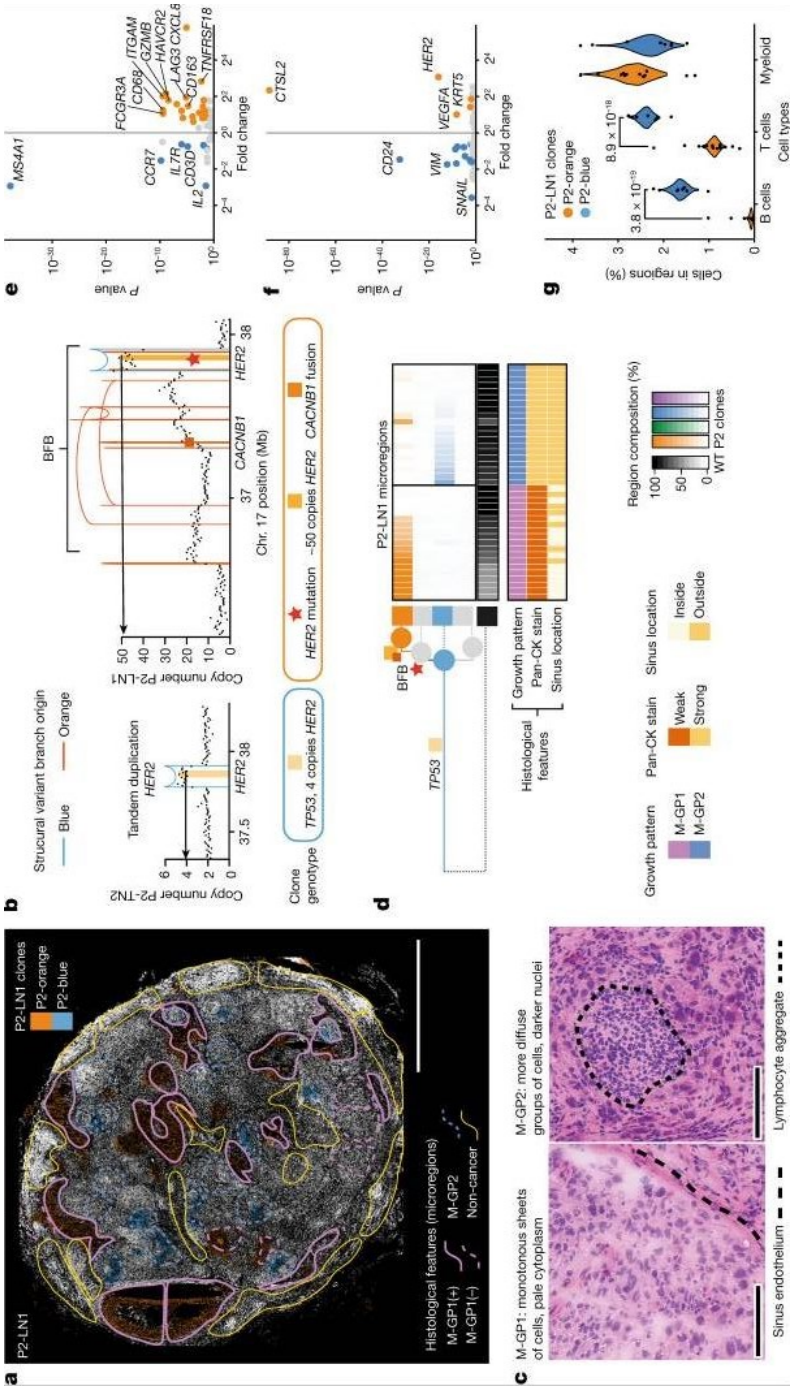


Figure 3.4: Spatial Subclonal Structure in Nodal Metastatic Disease. Reproduced from [41].

3.1.2 Insights offered by Spatial-omics

The above applications of spatial-omics can be achieved by offering insight into three characteristics of tumour tissue:

1. Cell composition.
2. Cell-to-cell interactions.
3. Ligand-receptor relationships between adjacent and distant cells.

The insights offered by spatial-omics are shown in Figure 3.5.

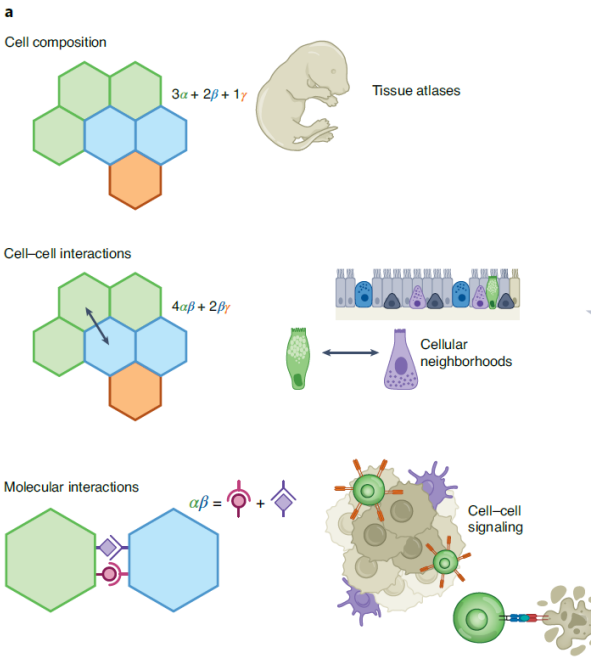


Figure 3.5: Biological Insights of Spatial-omics Reproduced from [42].

3.1.2.1 Spatial-omics Insights for Biomarker Opportunities

The potential therapeutic and biomarker opportunities offered by spatial-omics insights has been demonstrated in the work presented by Gaglia et al [43].

Cell proliferation is determined by the frequency of mitotic figures and percentage Ki-67 staining. It has diagnostic implications [44]. High grade tumours are associated with higher proliferation rates which is a well-recognised poor prognostic marker [44].

However, there are limitations with such an approach to assess tumour proliferation. Mitotic figures and Ki-67 index are more specifically associated with the mitotic phase of the cell cycle. Therefore, they cannot provide a global measure of progression through the cell cycle and consequently can underestimate the true proliferation rate [43]. Spatial-omics approaches were applied to address this question using multiplexed protein imaging on clinically-derived FFPE specimans across a diverse range of tumour types: breast, lung, colon and ovarian carcinomas, mesothelioma and gliomas [43].

A multivariate score of global proliferation was developed based on a balance between proliferation markers, Ki-67, PCNA, MCM2 and cell-cycle arrest markers, p21 and p27. The score is called the Multivariate Proliferation Index (MPI) [43]. A categorical scoring system is used:

MPI Score	Feature	Label
1	Positive balance of proliferation markers	Proliferative
-1	No expression of proliferation markers	Non-proliferative
0	Mixed balance of proliferation and arrest markers	Arrested

Table 3.1: MPI Scoring System. Reproduced from [43].

Each cell in a multiplexed tissue section was scored and compared with Ki-67. It was found that 39 - 72% of MPI+1 cells were Ki-67 negative (Figure 3.6) [43]. Therefore, Ki-67 scoring alone will underestimate the proliferation index of tumours in a clinical context.

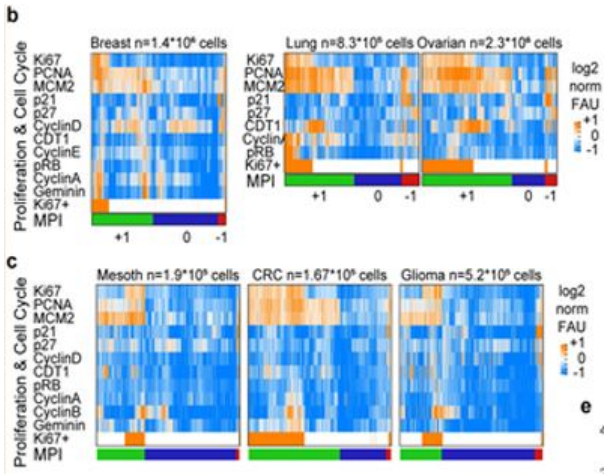


Figure 3.6: Heatmap comparing Ki-67 Positivity and MPI Score. Reproduced from [43].

Spatial maps of the MPI scores showed distinct spatial patterns (Figure 3.7a). Proliferating (MPI+1) and non-proliferating (MPI 0) states showed strong spatial self-state correlation (Figure 3.7b) [43]. In contrast, spatial cross-state correlation between proliferating and non-proliferating cells was weak (Figure 3.7b) [43]. Therefore, proliferating cells cluster together and away from non-proliferating cells.

Spatial self-correlation is well-fitted by a two-phase exponential decay model (Figure 3.7c) suggesting two levels of spatial physical structure [43]. Small niches between 10 - 30 μm are nested within larger structured neighborhoods spanning 100 - 300 μm (Figure 3.7d) [43].

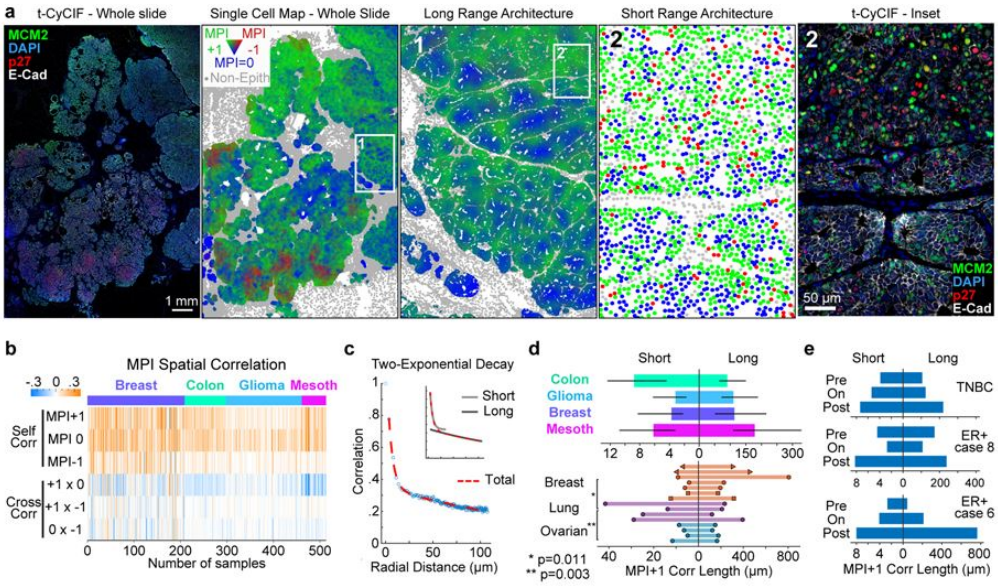


Figure 3.7: Spatial Maps of MPI Score and Correlation Metrics. Reproduced from [43].

To infer cell cycle dynamics, a new tool termed ccD-CMD (cell-cell distance and classical multidimensional scaling) was developed [43]. The core of the tool involves mapping the cell-cell correlation distance matrix onto a 2D multidimensional scaling for visualisation (Figure 3.8c - 3.8d) [43]. It is an interesting application of a domain-specified topographic constraint. Novel cell cycle coherence metrics, Inter-Octile Variation (IOV) and Circle Fit Distance (CFD) were integrated into the ccD-CMD tool to aid detailed evaluation (3.8e) [43].

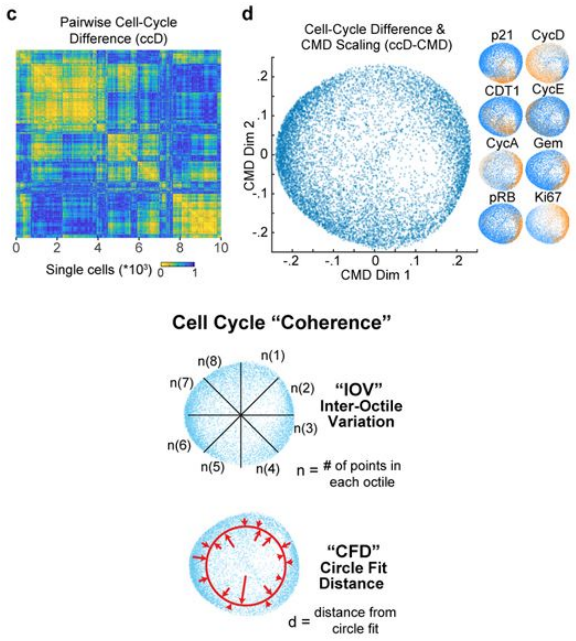


Figure 3.8: ccD-CMD and Cell Cycle Coherence Metrics. Reproduced from [43].

Galgia et al demonstrated it is possible to extract the cell cycle coherence metrics from multiplexed imaging in an interpretable manner (Figure 3.9a). MPI+1 cells from a cohort of 26 HER2 positive breast cancer samples (TMA1 and TMA2) were evaluated (Figure 3.9b). Samples displaying distinct metrics were manually inspected. Sample 1 is representative of an $IOV^{low}CFD^{low}$ representing cells displaying classical cell-cycle dynamics (Figure 3.9c) [43]. In contrast, sample 2 which is IOV^{high} is skewed to G1 phase and sample 3, CFD^{high}, represents a state outside currently recognised dynamics (Figure 3.9c) [43].

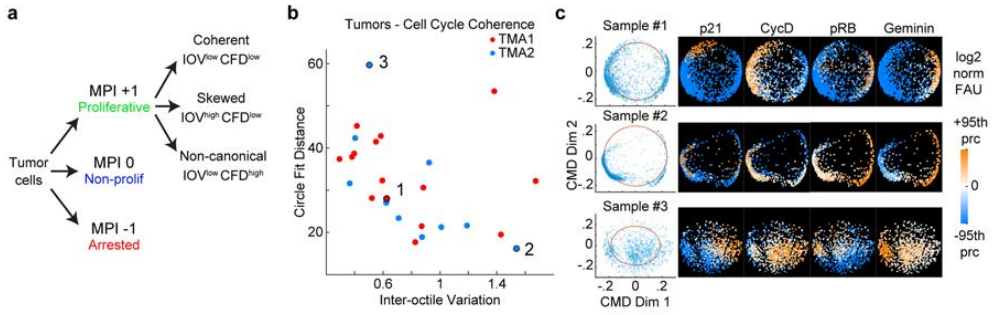


Figure 3.9: Cell Cycle Coherence in HER2 Positive Breast Cancer. Reproduced from [43].

Figure 3.10 demonstrates that MPI+1 fractions can be linked with annotated H&E sections [43]. Tumour subregions demonstrating distinct metrics across MPI classes can be ascribed to distinct histological regions, thereby suggesting these novel metrics summarise consistent and representative components of in-situ tumour behaviour [43].

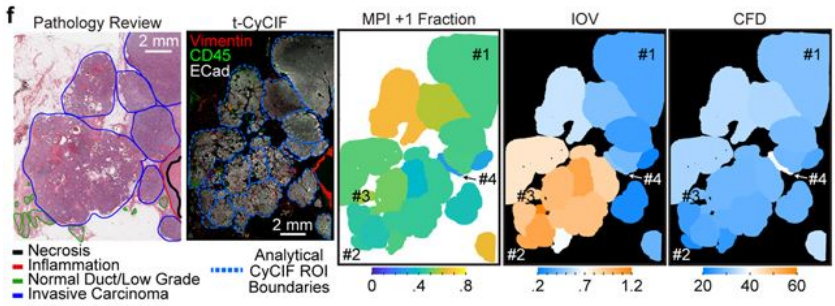


Figure 3.10: MPI Maps can be linked with Histology. Reproduced from[43].

A key component of the work was investigation of treatment-induced changes in the coherence metrics. Three clinical cases representing one case of TNBC and two cases of ER positive breast cancer were evaluated prior to starting treatment ('pre'), on treatment ('on') and after completion of treatment ('post') as depicted in Figure 3.11a [43]. I am particularly intrigued by the differences in MPI dynamics with treatment between ER positive and TNBC diseases.

ER positive breast cancer, a subtype with a broad range of treatment options, showed distinct changes in MPI proportions with treatment [43]. The changes were consistent between the two samples. Treatment induced a reduction in the MPI+1 and MPI-1 fraction with a clear rise in the MPI 0 fraction by the end of treatment (Figure 3.11h - 3.11i) [43].

In contrast, treatment induced little change in MPI dynamics for the case of TNBC disease (Figure 3.11b) [43]. It is recognised that with current treatment options, patients with TNBC are at greater risk of developing distant relapse compared to ER+ breast cancer. I wonder if these results are an early hint MPI dynamics could be explored as an ancillary predictive biomarker for treatment response. Clearly such dynamics would require evaluation across a much larger subset of breast cancer patients representative of the different molecular subtypes of breast cancers. It would be interesting to assess MPI dynamics in detail for HER2+ breast cancer.

The approach described above would lend itself very naturally for application in spatial transcriptomics (ST). Feature selection could be applied in a domain-specific manner, tailored to the biological process of interest. Gene expression could be easily assigned as 'on' or 'off' in a threshold-dependent manner. Genes with an expression level <1 would be assigned as 'off', genes with an expression level =>1 would be labelled as 'on'.

I propose these methods could be applied to offer new insight into the immunoregulatory landscape of human tumours. Tumour regions would be iden-

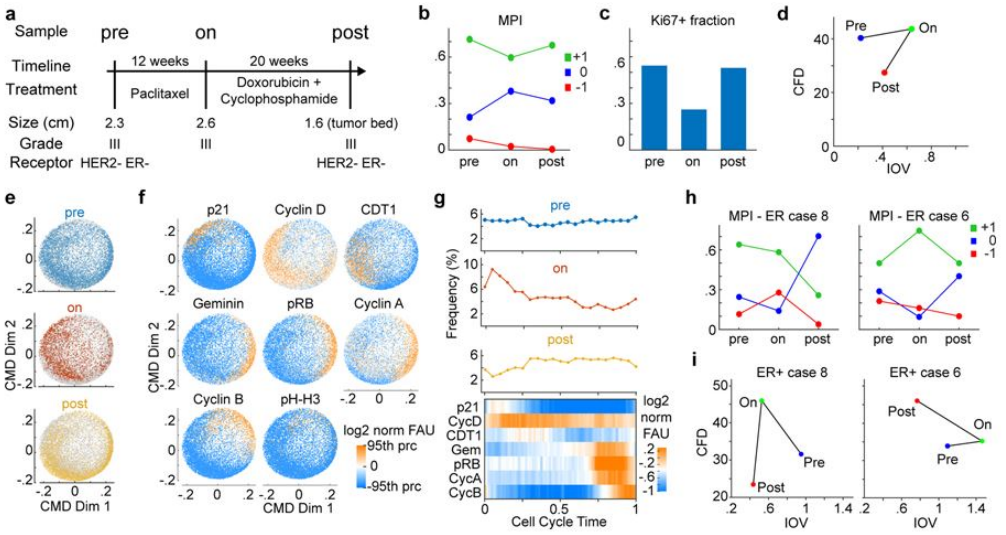


Figure 3.11: MPI Dynamics differ across Breast Cancer Subtypes. Reproduced from [43].

tified as immunostimulated, immunosuppressed or anergic using a composite immunoregulatory signature applied to ST data. An approach simmilar to ccD-CMD would be applied, except in this case, a simplex rather than a torus would be a more applicable topographic constraint to explore the dynamics between immunoregulatory transitions.

I foresee that these methods could be integrated into a clinical workflow in the longer-term. For T3/T4 tumours, grid-pattern biopsies could be extracted to develop a spatial map of tumour immunoregulation which could be used to guide clinical management. Immunosuppressive regions would be treated upfront with stereoablative radiation using the techniques currently applied for SABR/SRS followed by maintenance immune checkpoint inhibition. Such an approach would be particularly beneficial for tumour types associated with high surgical morbidity, eg GBM or pancreastic malignancies.

3.1.2.2 Spatial Transcriptomic and Proteomic Phenotyping

Spatial-omics broadly encompasses ST and multiplexed IHC technologies. ST offers the potential for insights into tumour biology through evaluation of gene

expression patterns. From a practical perspective, it is very expensive, cumbersome to perform and requires unique expertise and skills to be successfully implemented in clinical samples. Therefore, it is unlikely that ST will be applied in a broader and routine translational or clinical context in the medium or long-term future. In contrast, spatial phenotyping offered through multiplexed IHC may enable comparable biological insight to ST in a more practically achievable manner.

The routine delivery of multiplex IHC would be contingent on first addressing several unaddressed challenges. It is likely ST and multiplex IHC datasets share a common high-dimensional space. The integration of datasets and identification of such a space may require new computational tools which will require engineering to provide interpretable and biologically relevant insight. Identifying the shared latent representation will enable the spatial-omics field to map between technologies and circumvent some of the practical difficulties encountered when attempting to deliver new complex tools at scale. Questions also remain about how best to design and apply different multiplex panels across a range of biological contexts.

The translational scope of multiplex antibody panels has been demonstrated through the insightful, carefully designed work of Risom et al [45].

A retrospective cohort of 14 patients who presented with DCIS and later developed invasive breast cancer (IBC) after DCIS excision (termed 'progressors') were identified from the Washington University Resource of Archival Tissue (RAHBT) [45]. The cases were matched with 44 patients who did not proceed to later develop IBC (termed 'non-progressors). Normal tissue controls were included [45]. A 37-plex antibody panel covering tumour-intrinsic and microenvironmental markers was then investigated using the MIBI platform [45].

The study identified four cellular states, TME1 - TME4, spanning the transition from normal breast tissue to DCIS to IBC [45]. TME1 was enriched for markers associated with lipid metabolism in the normal breast, TME2 for mark-

ers associated with DCIS alongside increased myoepithelial proliferation and stromal CD4 T cells and mast cells and TME3 for IBC with markers associated with CAFs and collagen density [45]. TME3 was not enriched for tumour-specific markers. TME4 represented markers depleted in DCIS [45]. These states suggest that transition from in-situ to invasive disease requires the coordinated action of tumour and stromal cell types in a step-wise manner [45].

The most surprising result from the study was regarding the role of the myoepithelium in malignant transformation. The myoepithelium was found to be thinner and less continuous in the non-progressor cohort [45]. In contrast, the myoepithlium in progressors was found to more closely resemble normal breast tissue [45]. Closer scrutiny using gene set enrichment methods identified that ontologies relating to desmoplasia, stromal immune density and glycolysis were higher in non-progressors whilst ontologies relating to immunosuppression were higher in progressors [45]s. The current hypothesis is that breach of the myoepithelium facilitates immune cell and fibroblast activation in the stroma which facilitates immune recognition and control [45].

These findings can have wider clinical significance if supported in future additional studies. DCIS is a common clinical finding. There are two broad types of scenarios:

i. High grade DCIS alone or intermediate DCIS in the presence of additional higher risk pathological features such as comedonecrosis. In this setting, DCIS requires radical treatment with excision and adjuvant breast radiotherapy.

ii. Background DCIS is commonly found across all subtypes of invasive breast cancer. DCIS present at the surgical margin following breast conserving surgery requires further management with re-excision. This is commonly encountered in routine clinical practice, exposes patients to a second surgical procedure, delays starting adjuvant radiotherapy and adds further burden to the clinical workload.

The findings that myoepithelial depth may predict risk of malignant transfor-

mation could guide risk-adapted management of DCIS. Patients with a thinner myoepithelium may avoid the morbidity associated with a wide local excision and adjuvant therapy. Such patients may instead elect to proceed with close long-term monitoring. In contrast, patients with an intact myoepithelium would proceed with treatment as per the current clinical guidelines. Assessment of myoepithelial depth could be easily added to the existing breast pathology workflow.

I foresee that such spatially resolved findings could be expanded upon in a two-stage approach. Stage one would entail an international observational cohort study exploring the prevalence of myoepithelial disruption across a wide range of hospital settings covering both large academic centres and smaller district hospitals serving cultural diverse populations.

Patients would be subdivided into two groups depending upon myoepithelial depth and followed up to determine risk of relapse with invasive disease. The study would offer insight into the predictive value of myoepithelial depth and potentially shed greater mechanistic insight into such a surprising observation. Counterfactual modelling approaches could be explored to disentangle the risk of malignant transformation from the confounding role of existing treatments.

Stage two could then be explored if myoepithelial depth was identified to be a robust predictive feature for risk of future development of invasive disease. Stage two would be an interventional study based on myoepithelial depth.

Patients with thin myoepithelium would undergo close surveillance. Patients with thick myoepithelium would undergo current standard of care. It would be a practice-changing approach based directly on insights from spatial-omics techniques.

3.1.3 Opportunities in Data Interpretation

The opportunities offered by any new technology will be accompanied by new challenges in data processing and interpretation.

The practical constraints in conducting ST experiments together with the cost of the experiments results in low sample numbers per study. Typically, a ST experiment may consist of 10s of samples at most. However, the dimension size of the feature space, ie the number of genes, will be several fold higher. From an algebraic perspective, we are therefore dealing with an overdetermined system for which no exact solution may exist. The challenges are compounded by the paucity of power analysis tools for spatial-omics. Therefore, careful algorithmic selection is critical against such a backdrop. Low sample numbers also presents clinical limitations. It can be difficult to ascertain the clinical significance of results identified from a handful of patients.

Different ST technologies vary in their degree of multiplexing and the sensitivity/specificity of RNA capture which is ultimately reflected in the output data. Such systematic bias in data, particularly if under-recognised, can introduce artefactual anomalies into the results. These risks can be mitigated by recognising the differences between different technologies and introducing pan-technology QC metrics into the preprocessing steps.

ST data, in particular data from sequencing-based ST techniques, are very sparse. The sparsity observed is greater than that observed in dissociated single cell sequencing data. We are therefore working in the realm of a near-degenerate distribution since for most spatial locations, the expression of a gene will be zero.

The probability density function of a degenerate distribution is the Dirac delta function. The statistical properties of a degenerate distribution are not favourable. It requires advanced statistical tools and introduces further computational and algorithmic complexity. In general, a degenerate distribution

is best avoided.

3.1.4 Computational Tools

A summary of key processing tools is shown in Table 3.2. A deep dive into a key deep-learning ST tool, Tangram, is offered in Section 3.3.2.

Name	Algorithm	Function	Advantages	Disadvantages
RTCD [46]	Hierarchial poisson factorisation trained using Newton's method for optimisation.	Cell type decomposition	Leverages existing dissociated single cell datasets to accurately map cell types.	Computing the Hessian matrix is computationally expensive during training.
gimVI [47]	VAE maps spatial and single cell data into a shared latent space.	Gene imputation	Overcomes limitations with predetermined gene subset panels inherent to imaging-based ST.	Assumes cross-modality datasets share a common latent space which remains biologically informative under the constraints of unit gaussian regularisation.
SpaGCN [48]	Graph convolutional network: optimal graph structure identified through eigendecomposition of the graph Laplacian using first-order approximation methods for optimisation.	Integrates spatial gene expression with histology	Directly integrates ST and histology where distances between vertices remain biologically explainable.	Spatial gene expression is prioritised over underlying tissue architecture.
BayesSpace [49]	Bayesian model with Markov random field followed by clustering.	Cell type decomposition	Uses a fixed precision matrix across clusters which improves clustering stability.	Uses Gibbs sampling, a specific form of Metropolis-Hasting, which is computationally expensive. It is limited to spatial techniques where spatial spots are arranged in a lattice, e.g. Visium. Therefore, the technology itself imposes the neighborhood structure.
SpaOTsc [50]	Uses optimal transport to learn the mapping between non-spatial data and a small set of spatially-defined genes.	Spatial gene expression reconstruction	Defines a spatial metric for scRNA-seq data.	Requires scRNA-seq data matching the spatial sample of interest.

Table 3.2: Summary of ST Computational Tools

Key questions remain about the optimal approach to facilitate data interpretation. At present, there are few 'ready out of the box' tools tailored for the analysis of spatial-omics data. It is partially a consequence of the current novelty of the technology. Selected research groups have experience with spatial-omics technologies. Therefore, a lag in the development of data processing tools is to be expected.

Limitations currently remain at every stage of the data processing pipeline: data acquisition and management, data pre-processing and quality control and availability of algorithms which can reveal key biological insights from sparse high-dimensional data. To advance as a field, we will need to call on expertise from every discipline: deep learning combined with the theoretical rigor of Pure Mathematics, Information theory, Statistical Physics and Bayesian Statistics. It must be a truly inter-disciplinary effort.

The exciting work published by Galia et al [43] and Lomakin et al [41] gives us a glimpse into what is possible. The application of topographically-constrained multidimensional scaling ([43]) and gaussian process regression ([41]) is elegant. Existing tools with well-described properties have been repurposed for the spatial-omics domain combined together with a practical deep understanding of the underlying technology. This should be the guiding principle for how the field can develop, at least at this early stage.

I foresee two levels of processing tools: low-risk and high-risk tools. Low-risk approaches would include tools inspired from spatial statistics. Standard Dimensionality reduction tools, either PCA or autoencoders combined with a range of clustering approaches for which the most popular current approach is KNN-clustering, will likely form the backbone of this strategy. CNN-based segmentation algorithms, such as HoVer-Net [51], may be required for imaging-based spatial-omics. However, questions remains regarding the essentiality of segmentation in this setting. Segmentation-free approaches may gain traction

in the foreseeable future.

A less traditional approach would include tools such as Vision Transformers [52]. It is an intriguing possibility. Vision Transformers offer a key distinguishing property: the ability to model both short and long-range spatial dependencies. The nesting of spatial dependencies at different levels of physical scale is where spatial-omics technologies are uniquely placed to excel and offer unparalleled biological insight.

At this core, a Vision Transformer takes the convolutional feature map of the image together with the positional embedding of each image patch as input to the Transformer Encoder [52]. The Transformer uses a multi-head attention to then learn the short and long-range hierarchies within the image. The output embedding is then mapped back to the original feature map for real-world interpretation [52]. A schematic of the Vision Transformer is shown in Figure 3.12.

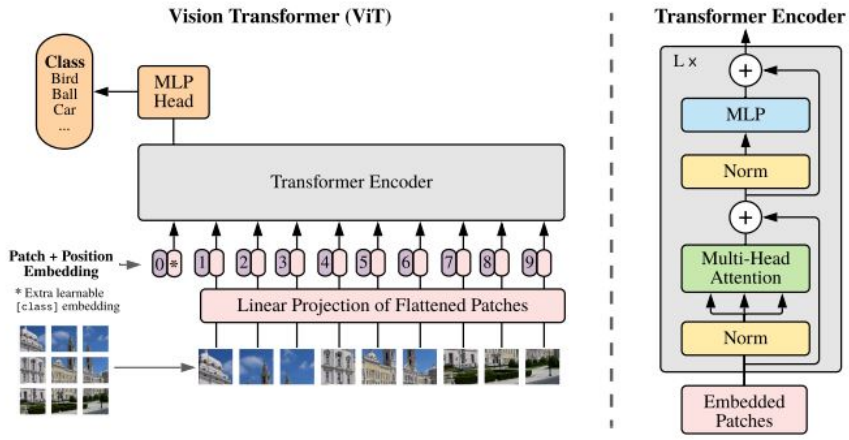


Figure 3.12: Architecture of the Vision Transformer.

This is cutting edge deep learning, requiring technically strong machine learning expertise and access to high-end compute. I anticipate training a Vision Transformer will be very challenging.

3.2 Methods and Results

3.2.1 Experimental Pipelines for Spatial Sequencing

Spatial transcriptomic technologies are bespoke platforms. Access to and experience in applying spatial transcriptomic technologies to biological samples are limited to selected research settings spread across North America and Europe.

A major component of the DPhil was in establishing novel experimental spatial sequencing pipelines and linking the newly formed pipelines with the collection of fresh clinical breast cancer tissue. Given the novelty of these technologies, exploration of spatial sequencing technologies was contingent upon such collaborations.

The two sequencing technologies of focus were:

i. Slide-seq

ii. CARTANA in-situ sequencing

In 2019, Slide-seq was newly published [34] and available solely through the Macosko or Regev lab at the Broad Institute of MIT and Harvard, Boston, USA. I therefore established a de novo collaboration with the Macosko lab, coordinated and negotiated the research proposal, agreed two-way financial obligations and implemented the inter-institutional legal agreement for the course of experimental work. I also arranged a Slide-seq training placement at the Broad Institute between November - December 2019. The successful completion of the placement enabled the release of a training set of pucks, the key spatially barcoded slides, to then set up the pipeline in Oxford according to best practice guidelines.

In parallel, a de novo commercial collaboration was established with CARTANA, a subsidiary of 10X Genomics. I negotiated the research proposal, the schedule of funding for the experiment and formalised the commercial-academic

agreement covering the use of human tissue in a commercial setting, sharing of data and publication rights.

The financial agreement differed in this commercial setting. CARTANA agreed to cover the cost of consumables and provided generous access to their in-house technical team during the pilot study. The general expectation was that a successful pilot would then form the basis for a larger grant encompassed within the wider research community at Oxford University.

3.2.2 Slide-seq: Experimental Pipeline

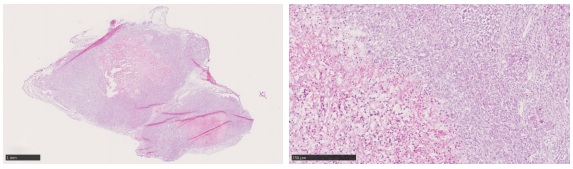
The Slide-seq pipeline was established using a test set of xenograft samples with the first set of pucks provided by the Macosk lab after completing the above training placement. The protocol is complex. It requires the coordination of multiple researchers and a range of equipment embedded across multiple institutions. Prospective experimental planning was critical.

Fully spatially profiled Slide-seq pucks are expensive to generate. The most expensive stage is the identification of the spatial location of each bead. It requires five daily consecutive sessions of 12-hour microscopy time. The set of pucks provided for training were produced in the same manner as a standard puck. However, the spatial profiling of beads, the most expensive step in the production process, is omitted. Therefore, the pipeline was set up in a cost-effective manner.

A set of seven fresh frozen xenograft samples from the Harris lab were accessed under the full direction of Dr. Bridges. The xenograft samples were generated from a previously completed experiment in which seven BALB/c nude mice were subcutaneously injected with the TNBC cell line MDA-MB-231 and treated with PBS.

Sample selection was based on H&E staining with representative images

demonstrated in Figure 3.13. The sample shown in Figure 3.13 was selected for downstream Slide-seq protocol optimisation since it exhibits high tumour content.



(a) **(b)**

Figure 3.13: H&E Assessment of Breast Cancer Xenografts Low and high resolution H&E images of the xenograft sample selected for optimisation of the Slide-seq protocol.

3.2.3 Slide-seq: Optimisation for Solid Tumours

The generation of the Slide-seq pipeline using xenograft samples revealed two findings:

i. The optimal conditions for sample sectioning in preparation for Slide-seq.

ii. The optimal protocol configurations for application of Slide-seq on solid tumour samples.

3.2.3.1 OCT embedding of Samples

Slide-seq is suitable for use on fresh frozen tissue. Standard practice for sectioning of fresh frozen tissue entails embedding tissue within OCT to provide a stable base from which sectioning is performed. The complete embedding of tissue within OCT is commonly adopted since it offers maximum tissue stabilisation to generate a full-face section.

Tissue contamination by OCT, which occurs as the cutting blade sweeps through the OCT block, does not interfere with standard histological stains. However, OCT interferes with RNA hybridisation onto the Slide-seq puck (Dr. Evan Murray, personal communication, 2019). RNA hydridisation onto the

puck is an early and essential part of the Slide-seq protocol. Unbiased RNA hybridisation is essential to achieving a comprehensive spatial transcriptomic assessment of the sample. A sample unsuitable for Slide-seq processing is shown in Figure 3.14.

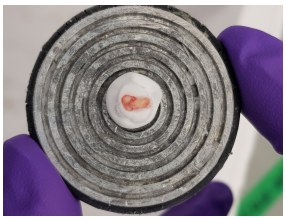


Figure 3.14: Slide-seq Sample Preparation. A clinical biopsy unsuitable for investigation by Slide-seq due to OCT tissue contamination.

Therefore, the amount of OCT used for tissue sectioning was kept to the minimium. A small amount of OCT is required to stabilise the biopsy sample onto the cryostat chuck prior to sectioning. This latter step required several cycles of iteration to identify the mininium effective OCT volume in a sample specific manner. Representative examples of samples suitably oriented for Slide-seq processing are shown in Figures 3.15a to 3.15c.

(a) **(b)** **(c)**

Figure 3.15: Sample Orientation during Slide-seq Tissue Sectioning. The avoidance of OCT contamination during sectioning is an important component in the Slide-seq protocol.

3.2.3.2 Slide-seq permeabilisation

Effective RNA hybridsation onto the spatial beads requires:

i. Sufficient tissue permeabilisation to allow for adequate RNA capture.

ii. Maintenance of adequate spatial resolution by avoiding excessive lateral diffusion of RNA.

The Slide-seq protocol has been published for application in murine brain tissue [34]. A schematic of the published protocol is shown in Figure 3.16.

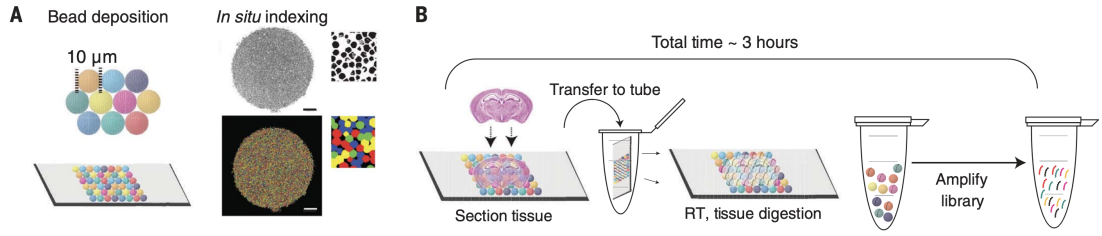


Figure 3.16: Slide-seq Protocol. Schematic of the Slide-seq protocol published for use on murine brain tissue [34].

Breast cancer tissue is different to murine brain tissue. Key histological differences in breast cancer tissue include marked cellular heterogeneity, variable fibroblast composition and disorganised spatial architecture. Such differences in tissue composition necessitate optimisation of the Slide-seq protocol in a tissue-specific manner.

In the protocol published using murine brain tissue, RNA hydridisation was performed for 15 minutes. For application on breast cancer tissue, a range of durations were explored for RNA hybridisation: 15, 30, 45, 60, 75, 90 and 105 minutes.

The quality of RNA hybridisation onto the spatially-profiled beads was assessed using the Agilent high sensitivity DNA kit prior to library preparation and analysed using the Agilent Bioanalyzer system. The Bioanalyzer results for the range of hybridsation conditions are shown in Figure 3.17.

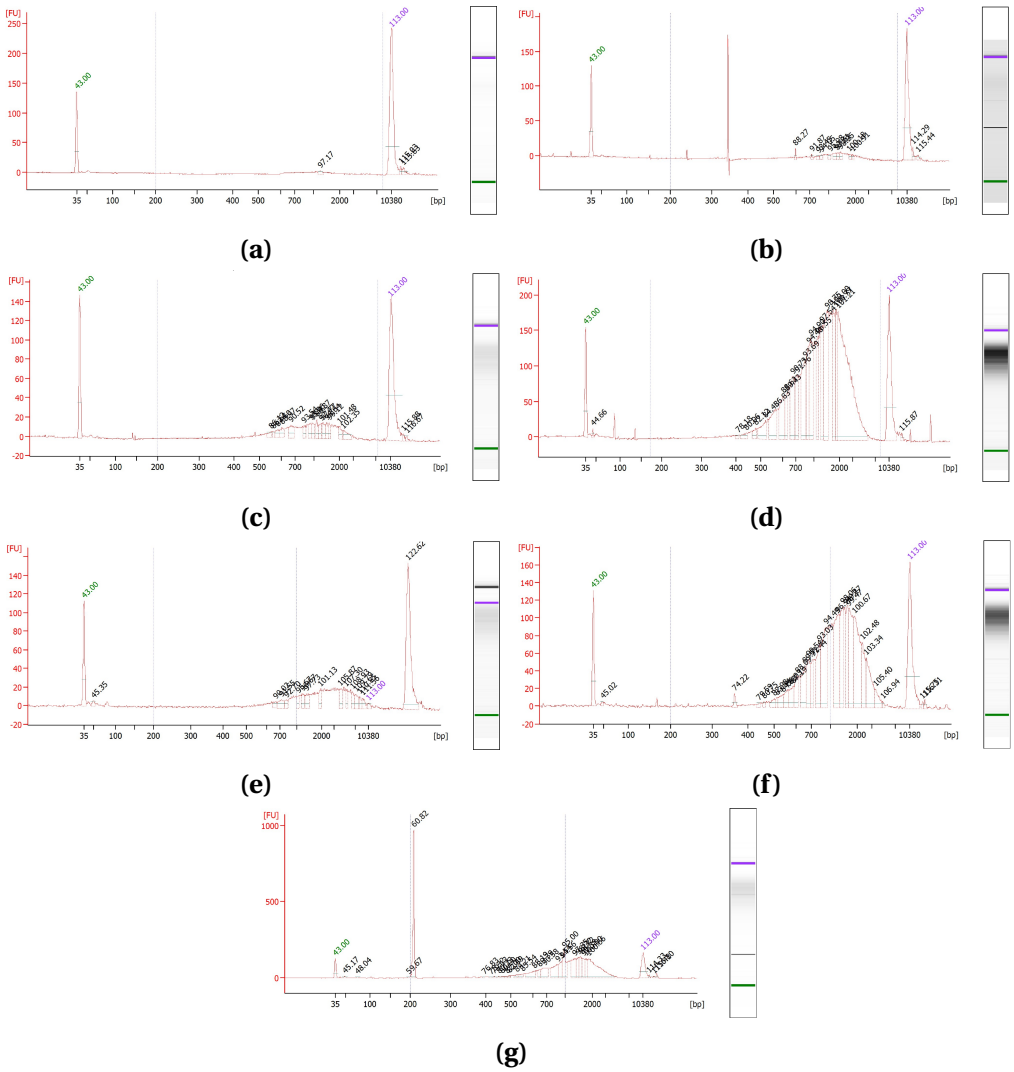


Figure 3.17: Optimisation of RNA Hybridisation for Slide-seq RNA hybridisation durations of 15 minutes (a), 30 minutes (b), 45 minutes (c, 60 minutes (d), 75 minutes (e), 90 minutes (f) and 105 minutes (g) were assessed.

The Bioanalyzer outputs a distribution of fragment lengths. The expected modal fragment length is 1000 - 2000 bp. The modal fragment length observed for the breast cancer xenograft sample was 1500 bp. The observed result is within the expected range. Optimal RNA output was obtained for a hybridisation duration of 60 minutes. Therefore, the Slide-seq protocol was modified to a RNA hydridisation duration of 60 minutes when applied to breast cancer tissue.

The experiment was completed under the close remote guidance of Dr. Evan Macosko, Broad Institute, USA.

3.2.4 Slide-seq: Application to Human Tissue

The successful establishment of the Slide-seq experimental pipeline in Oxford and optimisation of the protocol using xenograft tissues enabled the exploration of Slide-seq on clinical breast cancer samples.

Slide-seq was applied on patient-derived TNBC samples collected as part of the BRECO study. BRECO is a single centre prospective study investigating the relationship between breast cancer and the surrounding tissues. Human tissue, blood samples and clinical data are collected from patients undergoing primary breast surgery. Exclusion criteria include neoadjuvant chemotherapy or radiotherapy. All tissue was collected in accordance with BRECO REC requirements which are fully HTA compliant. The study REC reference is 19/SC/0025.

I regularly attended the breast MDT to identify potentially eligible study participants and liaised closely with the supporting clinical and research team. All samples were collected in a REC-compliant manner.

Sample collection was adjusted within the existing REC framework in order to optimise preservation of the spatial architecture of the samples. Samples were collected by snap freezing in liquid-nitrogen cooled 2-methylbutane. Such an approach was adopted because freshly collected tissue is warm. If warmed

tissue is placed directly into liquid nitrogen, it forms a gaseous vapour around the sample and may generate artefactual craacks within the tissue.

In contrast, 2-methylbutane has high thermal conducitivity [53]. Liquid-nitrogen cooled 2-methylbutane does not form a gaseous vapour around warm tissue. The tissue is quickly and evenly frozen [53], offering optimal preservation of the spatial architecture.

3.2.4.1 Slide-seq: Selecting the Optimal Clinical Biopsy

Biopsies were collected from three treatment-naive TNBC patients during the BRECO study. Per patient, five tumour and five normal tissue biopsies were collected.

Twenty tumour biopsies were assessed by H&E staining. The four best biopsies, as determined by tumour content, were selected for downstream investigation using Slide-seq.

H&E images of the biopsies selected for Slide-seq profiling are shown in Figures 3.18.

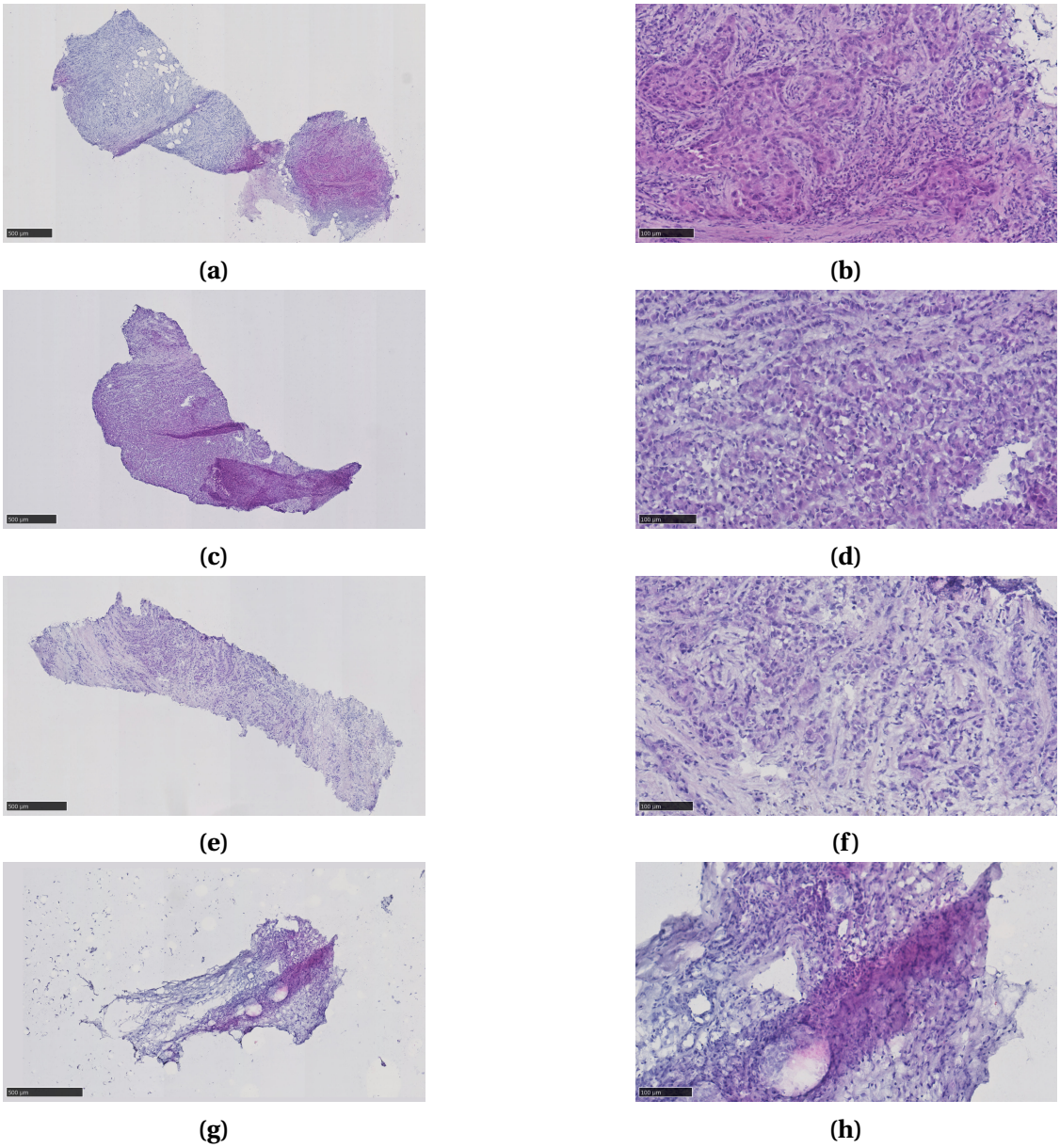


Figure 3.18: H&E Assessment of Clinical Breast Cancer Biopsies H&E assessment of breast cancer biopsies for Slide-seq profiling. Low and high resolution images of BRECO samples BC.18.3 (a and b), BC.20.2 (c and d), BC.20.4 (e and f) and BC.23.4 (g and h).

The biopsies processed using Slide-seq technology are histologically diverse. Approximate cell type proportions per biopsy are detailed in Table 3.3.

Biopsy ID	**Tumour (%)**	**Immune (%)**	**Fibroblast (%)**
BC.18.3	30	55	15
BC.20.2	80	10	10
BC.20.4	45	5	50
BC.23.4	30	10	60

Table 3.3: Cell Type Proportions per Breast Cancer Biopsy. Breast cancer biopsies selected for investigation under Slide-seq technology demonstrate wide variance in cell type proportions.

Sample BC.18.3 exhibits distinct and contrasting regions of infiltrative ductal carcinoma (Figure 3.19a) accompanied by peri-tumoural lymphocytic infiltrate with dense fibroblastic formation (Figure 3.19c).

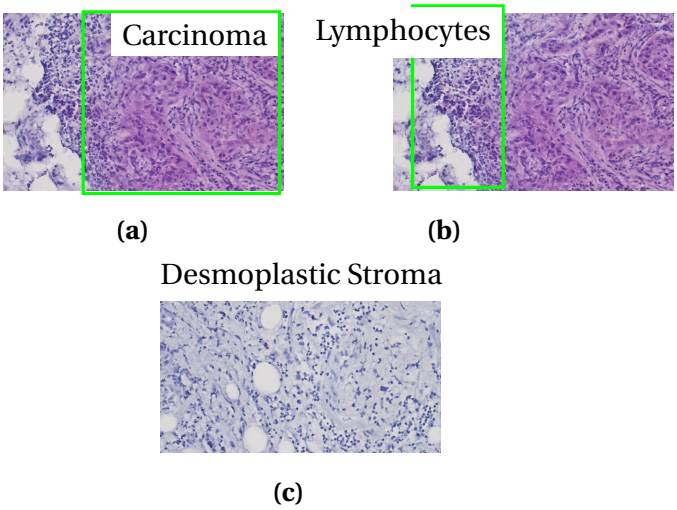
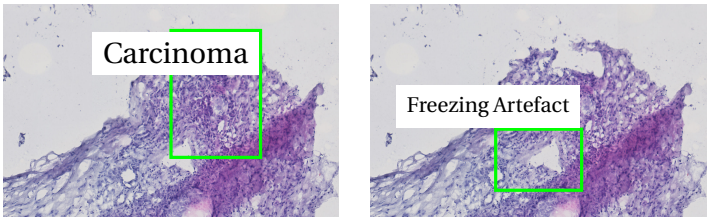


Figure 3.19: Desirable Histological Features.. Sample BC.18.3 exhibits well-defined regions of carcinoma (a), peri-tumoural lymphocytes (b) and desmoplastic stroma (c).

There are limitations with sample BC.23.4: the total tissue area is small, the tumour proportion is small (Figure 3.20a) and tissue artefacts (Figure 3.20b) were present.



(a) (b)

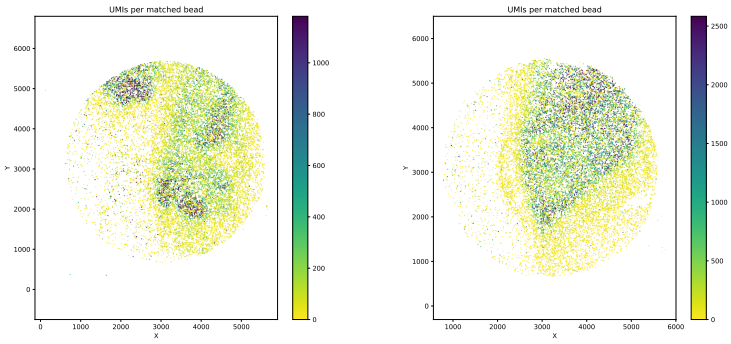
Figure 3.20: Less Desirable Histological Features.. H& E images of sample BC.23.4. Clinical biopsies may contain limited amount of tumour (a) and demonstrate freezing artefacts (b).

Biopsies BC.20.2 and BC.20.4 as shown in Figures 3.18c and 3.18e respectively were serial biopsies collected from the same patient. Sample BC.20.2 consists predominantly of a large region of infiltrative tumour demonstrating large nuclei and eosinophilic cytoplasm with some preservation of gross ductal architecture. Sample BC.20.4 contains a marked region of tumour infiltrate with a surrounding desmoplastic reaction and a paucity of resident lymphocytes. Biopsies BC.20.2 and BC.20.4 therefore exhibit representative spatial heterogeneity which can be present across the tumour of a single patient.

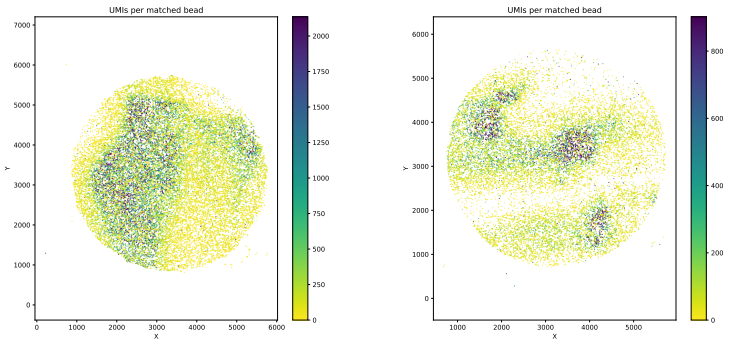
3.2.5 Slide-seq: Results

Four biopsies were selected because the Macosko lab donated four pucks as part of the agreed pilot study. Detailed information on the pucks, including the spatial map for each puck, is internal to the Macosko lab.

The modified Slide-seq protocol was successfully applied to four breast cancer biopsies. The spatial map of UMI counts on the pucks are shown in Figure 3.21.



(a) **(b)**



(c) **(d)**

Figure 3.21: UMI Spatial Map of Breast Cancer in Slide-seq Spatial maps were successfully generated for breast cancer samples BC.18.3 (a) , BC.20.2 (b), BC.20.4 (c) and BC.23.4 (d). *Figures provided courtesy of Dr. Evan Murray, Broad Institute, USA.*

The spatial UMI maps represent the UMI count per location, with yellow representing regions of a low UMI count and blue representing regions of a high UMI count. White regions do not contain tissue.

On inspection of the maps, variable regions of the pucks do not contain tissue. It is most evident for sample BC.20.2 (b). In contrast sample BC.20.4, a biopsy from the same patient but from a different spatial region, shows a more homogenous tissue map. These results suggest the presence of spatial heterogeneity in gene expression which can be detected at an intra-patient level. However, it is not possible to exclude additional technical biases from the presented results. The majority of spatial regions have a low UMI count which suggests additional computational challenges may arise due to a low signal-noise ratio. It is possible the low UMI regions have arisen due to a predominance of desmoplasia within the collected samples.

To further explore the question of tissue heterogeneity, I compared the spatial maps generated from my cohort of TNBC biopsies to a pre-print recently submitted by Hirz et al [54] in which samples from 19 patients with treatment naive prostate cancer together with paired normal tissue were analysed on the Slide-seq platform. Figure 3.22 shows four spatial maps generated from healthy prostate tissue, adjacent normal tissue in a patient with low-grade prostate cancer, low-grade prostate cancer and high-grade prostate cancer. The spatial maps were generated using the same Slide-seq protocol as used in my study.

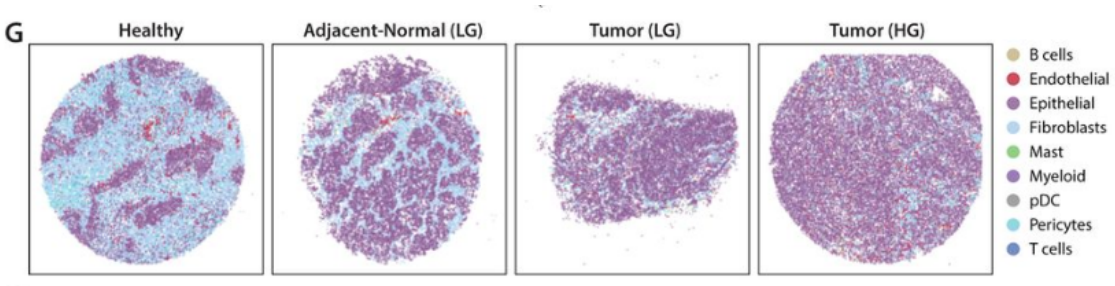


Figure 3.22: Cell type spatial map of normal and malignant prostate tissue. Spatial maps were generated on the Slide-seq platform. Figure reproduced from [54].

Comparison of the UMI spatial maps between TNBC and prostate cancer suggests differences in the spatial organisation of tumours arising from different anatomical regions. If the low UMI regions in the TNBC samples correspond to regions of desmoplasia, it suggests that desmoplasia is a more predominant feature of TNBC compared to prostate cancer. The spatial organisation of TNBC may be more akin to the organisation observed in normal prostate tissue.

3.2.6 CARTANA: Commercial Pipeline

The factors determining the optimal pairing of spatial transcriptomic technologies with tissue type are yet to be fully elucidated. Slide-seq offers an ex situ sequencing approach. In contrast, the CARTANA platform offers in situ sequencing, a conceptually different technique. One aim of the DPhil was to compare different spatial sequencing technologies to elucidate how best to apply different platforms in different research settings.

Eleven FFPE TNBC samples were selected from the Oxford Radcliffe Biobank (ORB) in close collaboration with Dr. Stephanie Jones. CARTANA spatial transcriptomic technology is suitable for FFPE or fresh frozen tissue. FFPE samples were selected for the CARTANA pilot study for three reasons:

i. Snap freezing offers an uncomplicated, time-efficient method for tissue preservation which does not alter the native structure of RNA and protein. However, it is not ideal for the preservation of tissue morphology, a key property under investigation as part of spatial transcriptomic studies [55].

The storage requirements for fresh frozen tissue is resource intensive. In contrast, FFPE preservation is preferable for maintaining spatial architecture with easier tissue storage requirements. This factor contributes to the widespread adoption of FFPE preservation in routine clinical diagnostic settings [55].

ii. Most clinical tumour collections house FFPE samples. Fresh frozen clinical

tumour collections are less common. A spatial transcriptomic platform which offers acceptable spatial transcriptomic assessment for clinical FFPE tumour samples would be invaluable to advancing our understanding of tumour spatial gene expression programs. The range of available tissue samples is much greater in the FFPE setting.

iii. The CARTANA technology was accessible via the CARTANA technical team situated between the USA and Sweden. FFPE tissue offers practical benefits when transporting clinical samples between countries.

All samples were collected from the ORB collection and together with Dr. Stephanie Jones, the ORB collection was screened by hormone receptor and HER2 status to identify TNBC samples in reverse chronological order. Recent samples were selected for investigation in preference to older samples. All samples were assessed by H&E, CA9 and GLUT1 IHC. CA9 and GLUT1 were selected because they are well-established markers of hypoxia [56] [57] [58].

Samples BBP-1467 and BBP-1557 were selected for spatial transcriptomic assessment on the CARTANA platform. IHC of samples BBP-1467 and BBP-1557 are shown in Figure 3.23.

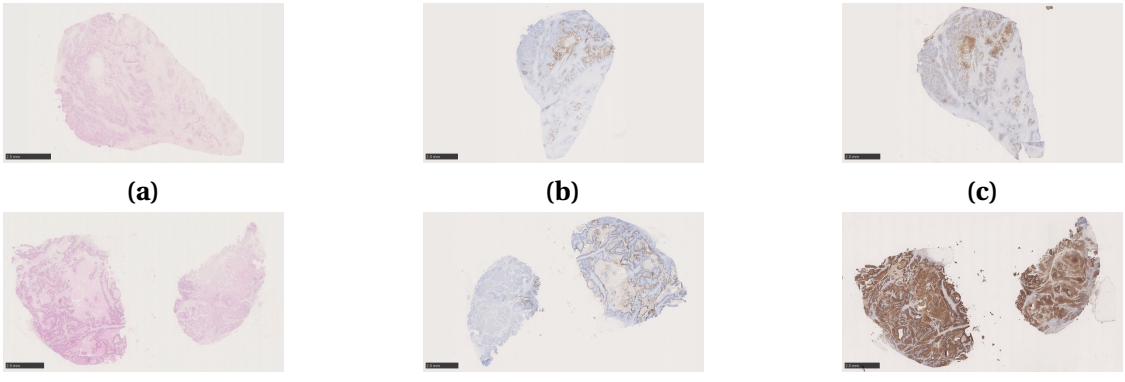


Figure 3.23: Histological Assessment of ORB Samples H&E (a), CA9 (b) and GLUT1 (c) of sample BBP-1467. H&E (d), CA9 (e) and GLUT1 (f) of sample BBP-1557.

These samples were selected because they exhibit well-circumscribed regions of hypoxia, as defined by CA9 positivity, with good tumour content and an

evident lymphocytic infiltrate. Furthermore, sample BBP-1557 demonstrates histological features suggestive of aggressive behaviour with nuclear atypia, extensive regions of necrosis and marked lymphocytic exclusion from the tumour compartment. A histological review of the samples was conducted together with Dr. Alistair Easton, a consultant pathologist, prior to sample selection.

3.2.7 CARTANA: Bespoke Spatial Hypoxia Panel

CARTANA generously created a hypoxia-specific panel tailored for the pilot study. The bespoke panel was optimised by the CARTANA in-house technical team and provided free-of-charge after negotiation. The panel is detailed in Appendix A.

An evidence based approach was adopted for target selection in the hypoxia specific panel using a combination of in-house unpublished and published [59] breast cancer single cell sequencing data. Genes which exhibited the following features in their expression distributions were included in the hypoxia panel:

i. Targets exhibiting a predominantly non-zero distribution pattern. A major challenge in single cell datasets is the commonly observed zero-inflated expression distribution.

ii. Targets exhibiting a non-uniform distribution profile to allow for the investigation of spatial heterogeneity.

All selected targets are associated with the transcriptional response to hypoxia [60]. Figure 3.24 demonstrates examples of the expression profiles of selected and deselected targets. *PGK1* (phosphoglycerate kinase 1) was selected and *PFKFB4* (6-Phosphofructo-2-Kinase/Fructose-2,6-Biphosphatase 4) was deselected.

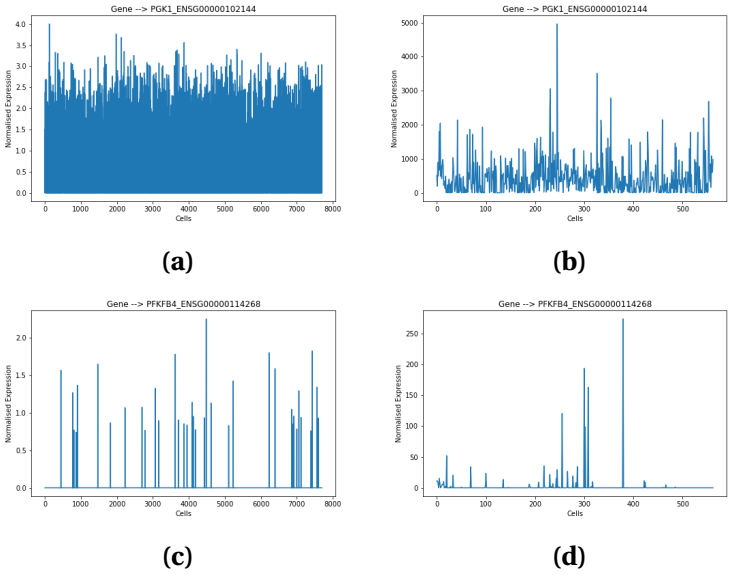


Figure 3.24
Single Cell Expression of Hypoxia-Associated Genes. Expression of *PGK1* in (a) BRECO sample BC.15 and (b) the single cell dataset published by Chung et al. Expression of *PFKFB4* in (c) BRECO sample BC.15 and (d) the single cell dataset published by Chung et al .

3.2.8 CARTANA: Results

CARTANA in situ sequencing was applied successfully to two archival FFPE breast cancer samples, ORB samples BBP-1467 and BBP-1556. Figure 3.25 demonstrates representative images of the transcriptomic expression of hypoxia targets in ORB sample BBP-1556.

From inspection, it is evident that hypoxia-related genes NDRG1, HK2 and BNIP3 (Figure 3.25(a) - (c)) demonstrate uniform and strong spatial expression across the sample, with the exception of the central region of the sample. It suggests that the tissue region analysed may represent an area of hypoxia. However, two additional markers of hypoxia, CA9 and PCAM1, show weaker spatial expression (Figure 3.25(d)). Different markers of hypoxia may undergo different mechanisms of spatial regulation. Alternatively, different levels of hypoxia (anoxia vs moderate hypoxia) may result in different spatial transcriptional responses.

The paucity of transcript expression in the centre may be due to technical factors related to sample sectioning and warrants future evaluation. No spatial clustering of hypoxia markers is evident by visual inspection of sample BBP-1556.

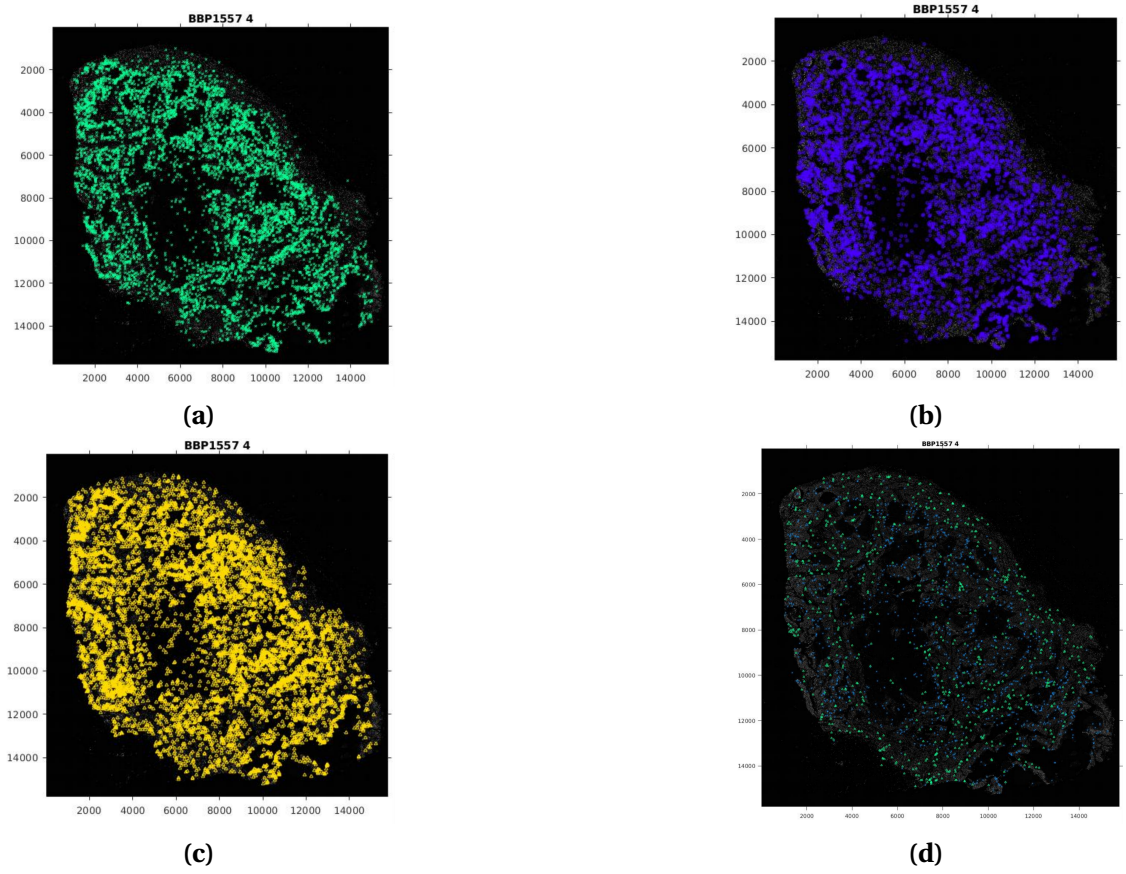


Figure 3.25: CARTANA Spatial Expression Profile of Hypoxia Markers. Spatial transcriptomic expression of (a) BNIP3, (b) HK2, (c) NDRG1 and (d) CA9 with PCAM1 evaluated by CARTANA in situ sequencing in one FFPE breast cancer sample.

The data output from the CARTANA in situ sequencing pilot study consists of two components:

i. A DAPI image outlining the position of each cell nucleus. By convention, the uppermost left-hand position is designated as position (0,0).

ii. The (x,y) coordinate position of each transcript.

The coordinate position of the nucleus is linked with transcript position to then proceed with downstream analysis. The first step in the analysis of a CARTANA spatial transcriptomic dataset is image segmentation of the DAPI

image to identify the coordinate position for each nuclear centroid. Correct algorithmic selection is critical for successful cell segmentation.

Threshold based image detection did not offer successful segmentation. A representative image of segmentation using thresholding is shown in Figure 3.26.

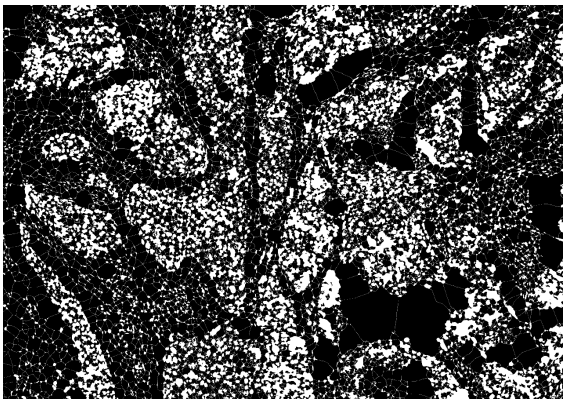


Figure 3.26: Cell Segmentation using the Watershed Transform Algorithm. Cell segmentation completed on ORB breast cancer sample BBP-1557 using ImageJ version 1.53i. The watershed transform algorithm binarises the pixel values of the image to identify regions of true foreground and background and then iteratively transforms the pixel values in regions of uncertainty in order to identify the boundaries between objects.

In contrast, image segmentation using deep learning approaches enabled successful image segmentation.

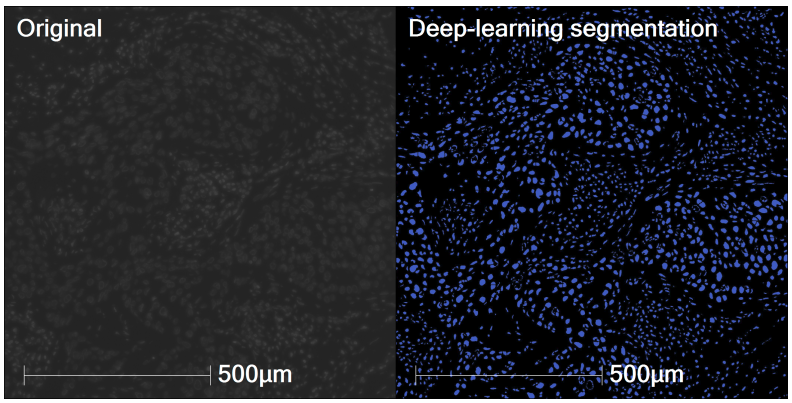


Figure 3.27: Deep-Learning Cell Segmentation. Application of deep-learning approaches for successful cell segmentation on ORB sample BBP-1557. *Image provided courtesy of Professor Viktor Kolzer, University of Zurich.*

The spatial transcriptomic map generated on the CARTANA platform differs to the spatial map generated by Slide-seq. It is likely some of the difference has arisen due to different tissue samples taken from different patients; different tissue processing techniques (fresh frozen tissue vs FFPE tissue); and due to differences in sequencing sensitivity. The greater sensitivity of CARTANA may be of value if the gene of interest, e.g. CA9, is known a priori to be lowly expressed. It is possible a gene such as CA9 may remain at undetectable levels by assessment on Slide-seq. The results signify the value of identifying the optimal spatial transcriptomic platform for the specific hypothesis under investigation.

3.3 Challenges and Limitations

The application of spatial transcriptomics to a diverse range of tissues and disease states represents an exciting frontier between discovery biology and technology development. It is a rapidly evolving field which has the potential to develop into a mainstream technique in the forthcoming years.

Important unaddressed challenges remain. These challenges include:

i. Experimental and computational QC of existing and emerging technologies.

ii. The need for new computational tools.

3.3.1 Quality Control

The two key quality control (QC) metrics for any technology platform is sensitivity and specificity. Optimisation of quality metrics for existing and emerging spatial transcriptomic technologies requires due consideration of two components:

i. The experimental workflow which can be further subdivided into the selection of the tissue and technology platform.

ii. Data management, pre- and post-processing analysis.

3.3.1.1 QC of the Experimental Workflow

The field of spatial transcriptomics is at an early stage in development. It is therefore extremely important early formative experiments are carefully designed which validate technological claims and allow for comparison across different technologies.

One step towards this aim would be the construction of standardised quality-controlled experiments. Such experiments would provide accepted benchmarks for validating the sensitivity and specificity of technologies. I propose two complementary approaches:

i. Mixed species experiments.

ii. Tumour-centric reference tissue sets.

3.3.1.1.1 Mixed Species Experiments Mixed species experiments are widely accepted QC experiments which have been previously successfully applied for dissociated single cell sequencing experiments [61]. The same principles could be applied for ST.

In a manner analogous to a fresh frozen tissue microarray, punch biopsies from human and murine could be co-located within adjacent units of an array mould. The mixed species tissue can then be sectioned onto the spatially barcoded glass slide.

3.3.1.1.2 Tumour-centric reference tissue sets The identification and generation of ST reference tissue sets could be conducted as part of a larger multi-institutional initiative. The reference tissue set would have generic and biology-specific requirements.

Generically, reference tissues should be easy to access, available at high volume from commercial vendors and easy to section for the non-expert histologist.

Tumour-specific requirements for a ST reference set have not been fully elucidated and require careful consideration. The key feature of a reference tissue is that it has a well-defined architecture preserved across difference samples which are visible to the non-expert ST user.

To date, most ST work has focused in the domain of neuroscience. Brain tissue has well-preserved anatomical structures whose architecture is closely linked with function. In this setting, ocular tissue has been proposed as an ideal reference tissue [42].

Tumour tissue presents unique challenges. Loss of normal tissue architecture is a property intrinsic to tumours and common to all tumour types in varying degrees. Tumour samples will not have a structure analogous to the eye. There are two avenues by which this challenges can be addressed:

i. Modification in the design of spatial slides.

ii. Generation of a ST normal tissue reference set.

The design of spatially profiled slides can be modified to incorporate inbuilt positive and negative experimental controls. An area of the slide, for example the top left hand corner, could be re-engineered to include standardised positive and negative controls. A positive control could be synthetic RNA arranged in a linear fashion. A negative control would consist of a blank square on the spatial slide.

The area of the spatial slide designated as the control region would be protected by a removable film. The tissue under investigation would first be sectioned onto the glass slide after which the protective film is removed. This approach would help facilitate ease in tissue placement onto the slide whilst safeguarding against contamination of the control region.

The addition of the control region may decrease the total area available on the slide for the experiment. However, the benefit of improved QC justifies

such a modification. This approach would be suitable for sequencing based ST platforms, such as Slide-seq.

A foundational experiment for tumour-specific ST may first consist of ST investigation of normal tissue. Normal tissue counterparts of the most common tumour types, eg colon and lung cancer, would be investigated. It would be a large-scale initiative. Access to such a valuable collection of tissue would be restricted to certain settings, such as the GTEX initiative or within the HCA consortium.

The scientific benefit of starting first with normal tissue is that it would enable systematic quantification of ST performance across different tissue types of varying composition across different technologies. Furthermore, normal tissue preserves spatial architecture more consistently across samples, unlike tumour tissue. It would be one pivotal step towards elucidating the optimal pairing between tissue type and technologies. ST experiments are expensive which will remain for the foreseeable future, compounded by macroeconomic inflationary pressures. Therefore, optimal tissue-technology mapping will help maximise the translational relevance and output of ST experiments and galvinize the field.

The insight gained through normal tissue ST will inform the design of tumour ST which may then shape how the field develops. Tumour samples collected prospectively for ST can be collected together with adjacent normal tissue from the same patient. The normal tissue ST results can be benchmarked relative to the reference normal tissue ST collection as a measure of experimental quality. If the sample is deemed acceptable, the normal tissue sample may then also form a patient-specific baseline from which tumour spatial heterogeneity can be formally quantified.

The collection of clinical tissue in a hospital setting is fraught with additional challenges and unexpected difficult-to-control delays. Such challenges can impact on the quality of collected tissue with downstream impacts on ST data quality. Therefore, a standardised QC metric applicable across hospital sites

over time is of benefit.

The limitation remains that normal tissue collected adjacent to a tumour site may not fully recapitulate the biology of normal tissue present in a non-cancer patient. However, the above suggested approach is intended to be pragmatic.

3.3.1.2 QC of Data Processing

A common bottleneck seen across many scientific disciplines are challenges in large scale data management and access to high performance compute. ST is a new field. The number of available data sets is small and the number of data processing tools relatively small. However, it is expanding rapidly. This offers an unique opportunity to standardise and automate these practically critical components whilst it is still possible to do so with comparative ease.

I propose to focus on three elements:

i. Base programming langugage for software development.

ii. Data formats.

iii. Image-processing pipelines.

Given the wealth of machine learning resources available for the Python language, I advocate the universal adoption of Python 3.8 or above for all ST-related data processing tools.

Some progress has been made towards standardising ST data formats. As one example, the Starfish initiative enables spatial data to be stored as a single object containing a tensor of pixel level data together with a standardised json file of key metadata and per file information in a technology agnostic manner [62].

At present, there is no single optimal solution. Starfish may offer a good starting point which can be further developed and modified as part of a wider discussion within the ST community. Alongside harmonisation of data formats, standardised primary ST datasets can be identified analogous to the 3K PBMC

dataset for single cell sequencing.

The need for standardised data formats accompanies the need for standardised image processing pipelines. At present, the processing pipelines are boutique, limited to research settings with established ST expertise and difficult to access for the ST beginner.

A major stride forward would the development of a central software repository accessible within an open access conda environment on AWS. Core processing tools recognised as critical for ST analysis by the wider community would be accessible free of charge for analysis of small datasets. Larger scale datasets or use of tailored compute intense tools would be chargeable as per the standard AWS pricing schedule. Such an approach would bring ST analysis aligned within the wider move into collaborative cloud-based computation.

3.3.2 New Computational Tools

ST enables investigation into the spatial dependency of gene expression. New computational tools are required to account for the complex underlying correlation structures inherent in such datasets to identify statistically and functionally relevant spatial domains. A plethera of new techniques are emerging. One promising technique is Tangram. A schematic of the Tangram algorithm is shown in Figure 3.28.

The Tangram algorithm learns a probabilistic map between dissociated single cell sequencing and ST data to construct a pan-genome single cell map of spatial gene expression. The dissociated single cell and ST data arise from the same tissue type and must share a common feature subspace. They are not required to originate from the same patient sample.

Gene expression from dissociated single cells are first randomly allocated onto the physical spatial locations represented by ST. The algorithm then opti-

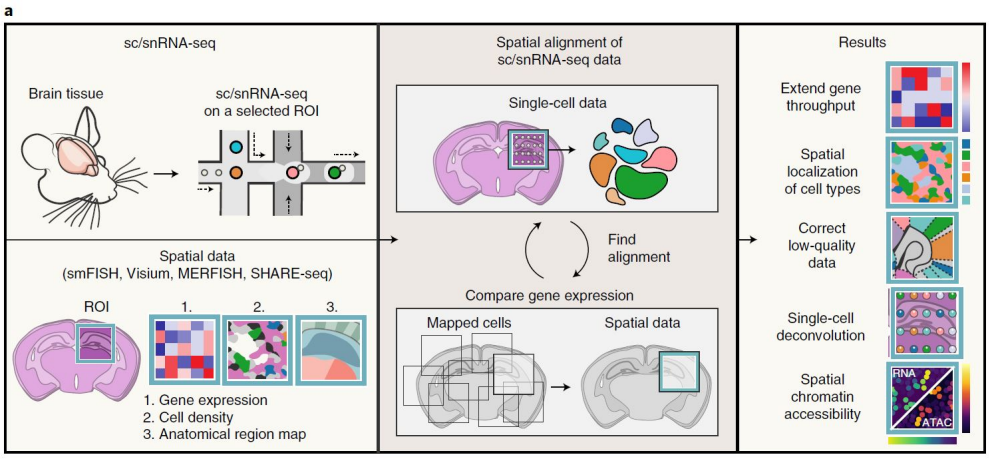


Figure 3.28: Overview of the Tangram Algorithm Reproduced from [63].

mises a non-convex objective to maximise the spatial correlation of genes.

The optimisation objective is shown in Figure 3.29. It aims to maximise the spatial correlation between genes. During model training, the objective is minimised using non-convex gradient-based optimisation.

$$\Phi(\widetilde{M}) = KL(\vec{m},\vec{d}) - \sum_{k}^{n_{genes}} cos_{sim}((M^T S)_{*,k}, G_{*,k}) - \sum_{j}^{n_{voxels}} cos_{sim}((M^T S)_{j,*}, G_{j,*})$$

Figure 3.29: Optimisation Objective of the Tangram Algorithm.

where M is the trained mapping matrix, KL is KL divergence, $\cos_{\text{sim}}$ is the cosine distance, k is the index for genes, j is the index for voxels, $M^T S$ is the spatial gene expression matrix predicted by M and G is the observed spatial gene expression matrix.

The objective consists of three terms which respectfully represent (a) Cell density (b) Gene expression over voxels (c) Gene expression per voxel. Voxel refers to the smallest resolution forming unit. As an example, a voxel in Slide-seq refers to the the 10 μm beads present on the puck.

For each of the above three terms, the predicted expression is trained to enforce similarity to the observed expression. Distance functions vary between

the terms. For cell density, 'distance' is measured using the KL divergence. In contrast, gene expression over voxels and per voxel uses cosine distance. In practice the second term, gene expression over voxels, is prioritised during model training.

Tangram is a useful tool. It may be constructive to benchmark future processing tools against Tangram.

3.3.3 Limitations

There are limitations with the presented body of work. The small sample size and sub-optimal histological quality of available samples limits the generalisability of deductions. Close and active collaborations with the primary technology developers, the Macosko lab and CARTANA, are required to interrogate and harness the full analytical potential of the generated datasets. Key metadata is available solely through collaborative engagement.

Slide-seq and CARTANA in-situ sequencing are not mainstream scientific techniques. They would be best explored in settings with established expertise in spatial sequencing technologies.

In summary, some progress has been made in establishing spatial sequencing pipelines for application in breast cancer. It has been demonstrated that Slide-seq and CARTANA in-situ sequencing can be applied to heterogenous breast cancer tissue. The work completed so far demonstrates that spatial sequencing approaches have the potential to offer deeper biological and translational insight of the tumour ecosystem.

4

A Single Cell Atlas of Breast Cancer

Contents

4.1 Introduction 92
4.2 Experimental and Bioinformatics Pipelines 92
4.2.1 Sample Collection 92
4.3 Results 93
4.3.1 Patient Specific Cell and Gene QC Thresholds 94
4.3.2 Critical Analysis of the Machine Learning Tools used in Single Cell Biology 97
4.3.3 Canonical Lineage Markers for Cell Type Assignment . . 104
4.3.4 Future Considerations 114
4.4 Discussion 120
4.4.1 Limitations 121
4.4.2 Future Directions 121

4.1 Introduction

The breast cancer ecosystem is characterised by homotypic and heterotypic interactions between cell types [64]. The emergence of single cell sequencing technologies is uniquely placed to both deconvolve the diverse interactions within the dynamic tumour system and offer the resolution required to develop and reshape the functional link between the molecular and clinical profiling of breast cancer [64].

The intention during the DPhil was to explore the TNBC ecosystem using droplet-based scRNA-seq technologies to provide a comprehensive single cell atlas of TNBC. It entailed the collection of clinical breast samples within a REC-compliant framework, the establishment of a new experimental pipeline for the processing of clinical breast cancer samples with scRNA-seq technologies and integration of the generated datasets with a newly established single cell sequencing bioinformatics pipeline.

It was anticipated that with such an approach, new features of the TNBC ecosystem could be identified in previously under-reported cell types. It would then be possible to link these features with clinically-relevant outcome parameters.

4.2 Experimental and Bioinformatics Pipelines

4.2.1 Sample Collection

Between January 2020 to March 2020, three patient samples at the Churchill hospital, Oxford were collected as part of the BRECO study.

BRECO is a single centre prospective study investigating the relationship between breast cancer and the surrounding tissues. Human tissue, blood sam-

ples and clinical data were collected from patients undergoing primary breast surgery. Exclusion criteria included previous chemotherapy or radiotherapy. Tissue was collected within a HTA-compliant framework. The BRECO REC reference is 19/SC/0025.

The clinical and histological characteristics of the recruited patients are detailed in Table 4.1.

Study ID	TNM Stage	Hormone Receptor Status
BC18	pT3 (50mm) pN1a (1/2) M0	ER 2/8 PR 0/8 HER2 negative
BC20	pT4 (110mm) pN3a (33/33) M1	ER 0/8 PR 0/8 HER2 negative
BC23	pT3 (55mm) pN1a (3/18) M0	ER 3/8 PR 0/8 HER2 negative

Table 4.1: Post-Surgical Pathological Characteristics of Breast Cancer Samples

Patients were recruited on the basis of ER, PR and HER2 immunohistochemistry from the diagnostic breast biopsy. All recruited patients were ER 0/8 PR 0/8 HER2 negative on the initial diagnostic biopsy. Receptor status can however vary when repeated on the definitive final surgical specimen. Table 4.1 details hormone receptor status from the final surgical specimen.

The sample and data processing pipelines are detailed in Sections 2.4.0.4 and 2.4.0.5. These pipelines were set-up as part of the DPhil.

4.3 Results

The skills for analysing single cell data were self-taught during the DPhil. I will present a summary of the generated results, a critical appraisal of the theoretical and practical framework underpinning the analytical approaches and potential avenues for future exploration.

4.3.1 Patient Specific Cell and Gene QC Thresholds

Analysis of scRNA-seq data involves multiple QC steps linked to gene and cell based parameters.

4.3.1.1 Gene QC

Genes expressed in minimum of 3 cells, a filtering threshold which is commonly adopted by the single cell community, were selected.

4.3.1.2 Cell QC

Cell QC is based on assessing number of counts per cell, number of features per cell and the percentage of counts originating from mitochondrial genes per cell [65]. During data preprocessing, the distribution of these three covariates were manually inspected. Thresholds for cell selection were identified in a sample-specific manner.

Thresholds were selected to achieve the following objectives:

i. Outlier cells with low counts or low feature levels were deselected. This step aims to exclude cells for which membrane damage has resulted in leakage of cytoplasmic mRNA [65].

ii. Outlier cells with high mitochondrial counts were excluded. Cytoplasmic mRNA leakage leaves mtRNA within the cell, resulting in high mitochondrial counts [65].

iii. Outlier cells with high counts or high feature levels were deselected to remove potential cell doublets [65].

The application of the above thresholds results in the construction of an unimodal covariate distribution. The threshold settings applied to each sample are detailed in Tables 4.2 to 4.4.

Study ID	Median Count	Lower Threshold	Upper Threshold	Median Absolute Deviation
BC18	3504	415	29609	2.5
BC20	3243	384	27365	1.9
BC23	1965	154	25068	2.95

Table 4.2: Thresholds for Total Counts

Study ID	Median Features	Lower Threshold	Upper Threshold	Median Absolute Deviation
BC18	1209	246	5934	2.2
BC20	1310	236	7262	2.25
BC23	830	166	4158	2.8

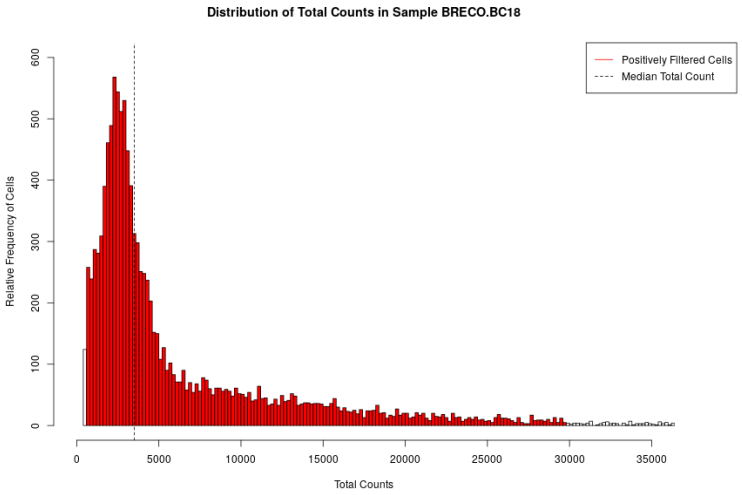
Table 4.3: Thresholds for Total Features

Study ID	Median Mitochondrial Content	Lower Threshold (%)	Upper Threshold (%)
BC18	7.19	0	20
BC20	5.02	0	20
BC23	4.77	0	20

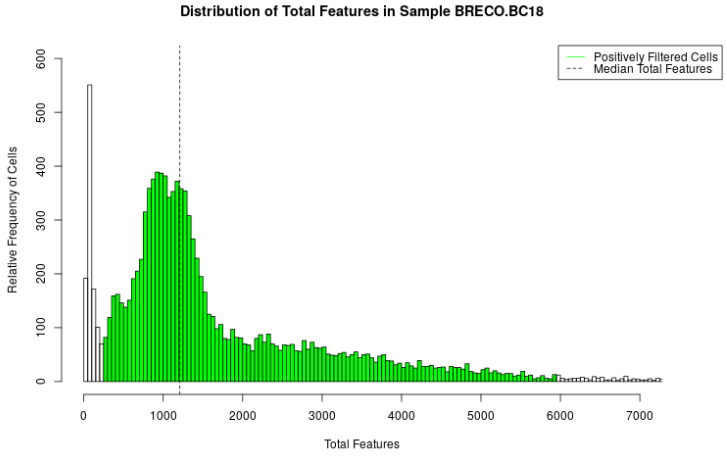
Table 4.4: Thresholds for Mitochondrial Counts

In the forthcoming sections, QC results for sample BC18 are presented. The presented results are representative of the QC results observed for samples BC20 and BC23.

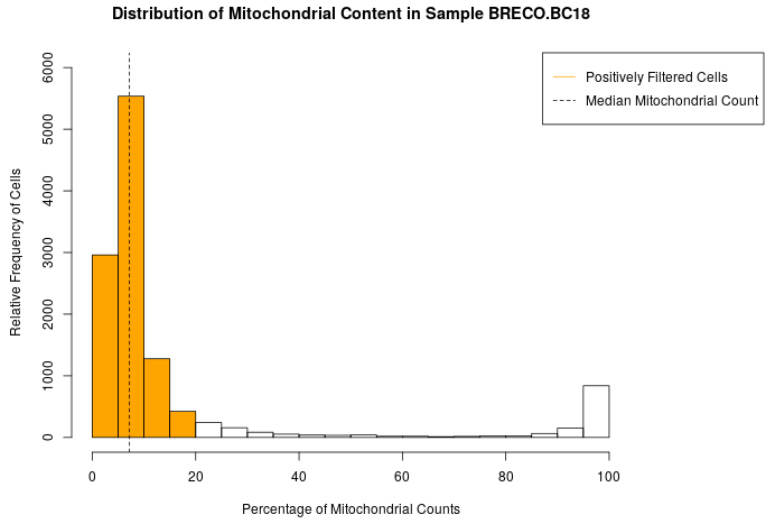
The covariate distributions evaluated for cell QC are shown in Figure 4.1. Positively selected cells are highlighted by the coloured histogram bins.



(a)



(b)



(c)

Figure 4.1: Sample BC18 Filtering Thresholds. Threshold-based preprocessing of (a) total counts, (b) total features and (c) mitochondrial percentage for sample BC18.

The final set of viable cells is shown in Figure 4.2. The selected viable cells form the data set used for all downstream analysis.

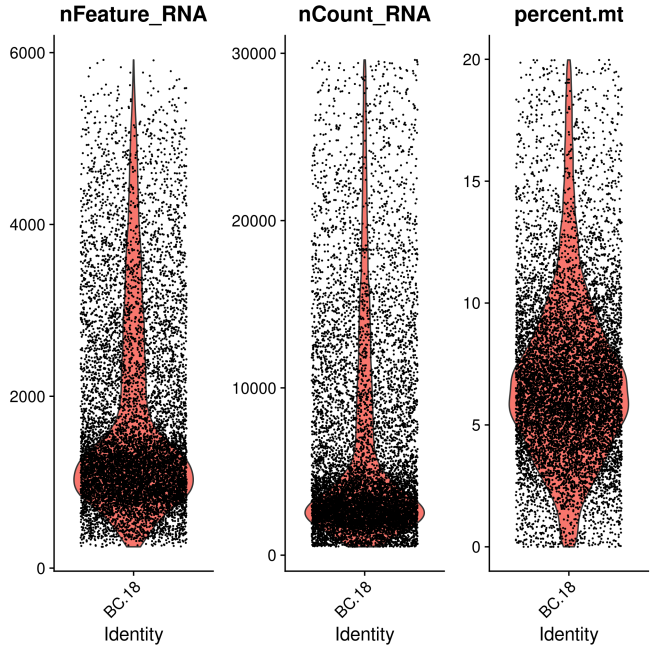


Figure 4.2: Viable Single Cell Selection. QC metrics for viable cells selected from BC18.

4.3.2 Critical Analysis of the Machine Learning Tools used in Single Cell Biology

Following completion of sample pre-processing, the scRNA-seq analysis pipeline consists of:

i. Linear dimensionality reduction

ii. Graph-based clustering

iii. Low-dimensional manifold visualisation

4.3.2.1 Linear Dimensionality Reduction

A range of dimensionality reduction methods are used in the scRNA-seq pipeline to introduce computational efficiency. A commonly used technique is Principal

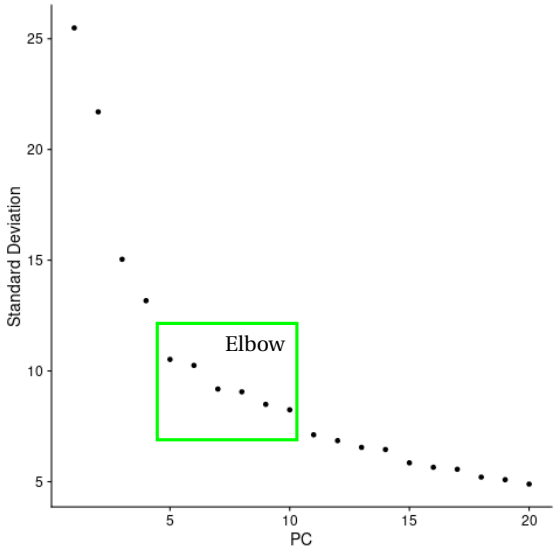
Component Analysis (PCA) [66].

PCA has its theoretical origins in Linear Algebra in which the gene expression per cell is compressed into a linear combination of the original feature space [67]. The original feature space is projected onto a lower dimensional space by using the eigenvectors of the covariance matrix of the original dataset [67]. The transformed variables, also known as principal components, are orthogonal transformations. Therefore, each principal component is uncorrelated in the new reduced feature space [67].

Key limitations of PCA include difficulty with interpretability. It can be challenging to ascribe practical significance to each principal component [68] [67]. PCA assumes a linear relationship between the features in the original domain space. The assumption of linearity may not always hold in biological systems [67]. Therefore, alternative approaches such as autoencoder architectures can be explored to enable non-linear transformations to a lower dimensional latent space.

scRNA-seq analysis involves identifying the total number of principal components required to capture biologically relevant variance in the dataset [65]. Therefore, the selected number of principal components is a hyperparameter.

The scree plot is a commonly used graphical tool for principal component selection [68]. It demonstrates the magnitude of variance contributed by each principal component. Principal component selection is guided by the elbow of the scree plot (Figure 4.3).



(a)

Figure 4.3: Scree Plot. The number of principal components selected for downstream analysis is determined by the location of the elbow.

The number of principal components required to capture biologically relevant information independently for each sample was manually identified. The scree plot and principal components selected for BC18 is shown in Figure 4.4.

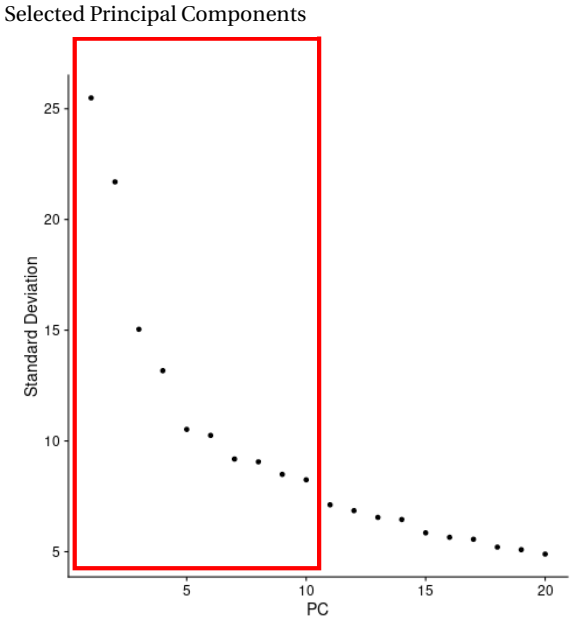


Figure 4.4: Principal Component Selection. Ten principal components were selected for sample BC18.

4.3.2.2 Graph-Based Clustering

The next step in scRNA-seq analysis involves graph-based clustering [69]. Graph based analysis is used across a broad range of disciplines [69]. They can expressively capture complex relationships across individual cells [69].

The cells are embedded into a K-nearest neighbour (KNN) graph, a commonly used directed graph structure. In a KNN graph, each cell is represented as a node. Cells with similar expression patterns are linked by an edge [70] [71]. A range of metrics exist to determine the distance between the expression patterns of cells [72].

In the analysis of this dataset, euclidean distance was used to calculate distance between gene expression in the PCA space. Euclidean distance is the most commonly used metric in the majority of scientific disciplines [72].

The Euclidean distance is defined as [73]:

$$d(x,y) = \sqrt{\sum_{i=1}^{n}(x_i - x_j)^2}$$

Exploration of metrics and the geometric implications on the KNN graph lies outside the scope of the thesis discussion.

Cells are then subsequently clustered using modularity optimisation methods [74]. Modularity measures the structure of networks. A high modularity score reflects a highly connected network [74]. In contrast, a low modularity score reflects a sparsely connected network [74]. The Louvain hierarchial clustering algorithm was used to perform unsupervised clustering on each sample by iteratively grouping cells to optimise the modularity of the network [75].

4.3.2.3 Low-Dimensional Manifold Visualisation

In scRNA-seq analysis, graph based clustering is followed by visualisation and interpretation of the data in a low-dimensional manifold.

Uniform Manifold Approximation and Projection (UMAP) is a commonly used technique in single cell analysis [76]. It is a relatively new algorithm, first published by McInnes et al in 2018 [77]. I will proceed with a critical appraisal of the UMAP algorithm. The proposed justification for its application in the analysis of scRNA-seq data will also be explored.

The UMAP algorithm has its roots in manifold learning and topological data analysis. It aims to capture the true topological structure of the data and learn the optimal low-dimensional topological representation [77]. A rigorous theoretical perspective on the algorithm requires due consideration of:
i. Simplicial complexes and their application to a finite set of data points.
ii. Manifold learning and the required assumptions when applied to real-world data.
iii. Learning the optimal low-dimensional representation.

4.3.2.3.1 Simplicial Complexes Continuous topological spaces are complex [78]. Therefore to facilitate their tractable exploration, topological spaces can be decomposed to their basic building block, simplices [78]. A simplex is a high-dimensional generalisation of a triangle. They play a critical role in algebraic topology and provide a geometric way to build a k-dimensional object [78].

A k-dimensional simplex is built by taking the convex hull of k+1 independent points [78], eg 0-simplex is a point and 1-simplex is a line. In this way, a topological space can be defined.

In order to apply the tools in topology to a set of finite data points (eg gene expression from a set of single cells), the open cover of a topological space must

be defined in order to construct the simplicial complex.

An open cover is a family of sets whose union is the whole space [79]. The simplicial complex is generated by the intersection of non-empty sets [78]. The Nerve theorem provides the theoretical guarantees that the intersection of the sets represents the underlying topological space in a meaningful way [80]. A detailed exploration of the Nerve theorem lies outside the scope of the thesis.

By assuming that the data points lie on a metric space (ie we can measure distances between points), an open cover of the data points can be generated by creating balls of fixed radius about each data point which can then learn the topology of the space [78]. The true underlying topological space cannot be directly accessed. Therefore, one can only expect to learn a reasonably good approximation of the true space [78].

In practice, the space is represented primarily by 0-simplices (points) and 1-simplices (lines). It constructs a representation of the space in the form of a graph which can be accessed in a computationally tractable manner.

4.3.2.3.2 Manifolds of Real-World Data The application of the above theoretical principles can be challenging with real-world data. The key challenge relates to the distribution of data on the manifold [77].

Real-world data is rarely uniformly distributed on its manifold [77]. The topological space is warped across the manifold according to the density of the data distribution. Therefore, identifying the radius of the ball around each data point is challenging [77]. Identification of the ball radius around each data point is required to construct the open cover [77].

By applying some standard Riemannian geometry, a locally varying metric which is equivalent to a ball of unit radius can be generated around each data point, thereby in effect estimating the Riemannian metric of the manifold [77]. A detailed discussion on the Riemannian geometry of manifolds lies outside

the scope of the thesis.

There are additional challenges when applying manifold learning to real-world data which can impact on the tractability of learning. These challenges include, but are not limited to, maintaining the property of local connectivity and addressing incompatible local metrics [77].

4.3.2.3.3 Learning the Low-Dimensional Representation By this stage, a topological representation of the data has been constructed. The next step involves identifying a low-dimensional representation which has two properties:

i. It preserves similar topological structure to the high-dimensional representation.

ii. It is a good representation.

Low and high dimensional data will lie on different manifolds. By explicitly defining distance in both manifolds to be Euclidean distance, preservation of structure between different manifolds can be partially achieved [77].

The question of identifying a good low-dimensional representation in essence involves identifying the closest match to the original high-dimensional manifold [77]. This therefore becomes a standard optimisation problem.

Each data point in the manifold is represented as a vector of probabilities for a range of k-dimensional simplices. Since a simplex can either exist or not, these are Bernoulli random variables. Therefore, cross-entropy loss is the ideal optimisation objective. The optimal low-dimensional representation is learnt by minimising the cross-entropy loss (Figure 4.5) using gradient descent optimisation methods [77].

UMAP thereby provides one potential mathematical framework for identifying the optimal low-dimensional embedding of scRNA-seq data. The low-dimensional embedding enables the efficient, scalable and interpretable anal-

$$\sum_{e \in E} w_h(e) log\left(\frac{w_h(e)}{w_l(e)}\right) + (1 - w_h(e)) log\left(\frac{1 - w_h(e)}{1 - w_l(e)}\right) \tag{4.1}$$

Figure 4.5: Cross Entropy Loss Function

ysis of the data [77] [76].

The generated UMAP plots for the three breast cancer samples collected as part of the BRECO study are shown in Figure 4.6. The UMAP plots exhibit intra- and inter-patient heterogeneity at single cell resolution for the breast cancer samples.

Each point on the plot represents an individual single cell. Each cluster represents a different cell type. The attribution of cell type identity to each cluster is described in detail in Section 4.3.3.

Figure 4.6 shows that the samples exhibit intra-patient heterogeneity with a distinct number of cell types per patient. Inter-patient heterogeneity is demonstrated by the varying number of clusters between patient samples and a differing number of cells per cluster between patients.

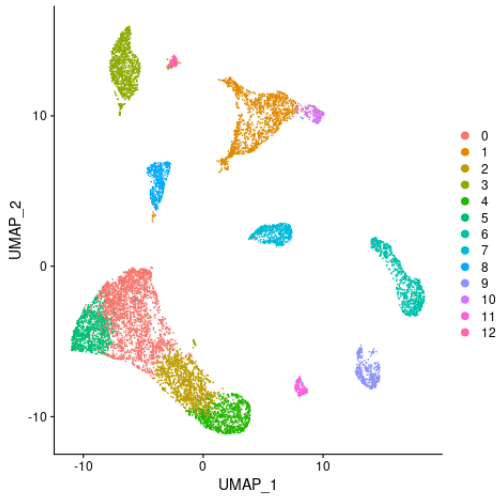
In summary, UMAP plots of single cell gene expression in breast cancer samples may allow identification of features of heterogeneity which have been previously described to correlate with treatment response and overall survival outcomes in patients [81] [82] [83].

4.3.3 Canonical Lineage Markers for Cell Type Assignment

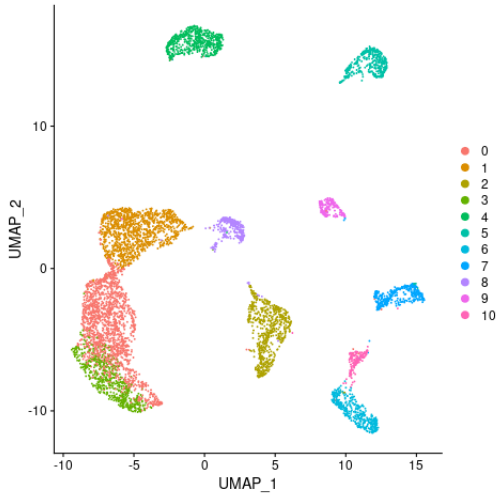
The next step in scRNA-seq analysis involves assignment of cell type identity to the generated clusters using unsupervised learning approaches [65]. Cell identity can be performed using two broad techniques:

i. Manual use of canonical lineage-specific marker sets [84].

ii. Automatic cell type identification using single cell references [84].



(a)



(b)

Figure 4.6: Single Cell 2D UMAP Projections. UMAP plots for sample (a) BC18 and (b) BC23. The legend details the cluster ID assigned by the KNN clustering algorithm to each cell. It is an unsupervised clustering algorithm. The total number of clusters is a predefined hyperparameter.

Manual annotation requires the use of gold standard markers associated with cell type, eg CD3 is an universally accepted marker of T cells [65] [84]. Marker sets were adopted that are associated with the major cell types previously identified and published in breast cancer [64] [9] [85].

The markers associated with each cell type are detailed in Tables 4.5, 4.6 and 4.7. The marker sets were kindly shared by Dr. Thomas Carroll. The same markers

have been successfully adopted in the single cell investigation of patient-derived oesophageal tumour samples.

Epithelium	Endocrine	Endothelium	Fibroblast
EPCAM	ERBB2	VWF	PDGFRA
MUC1	AR	CD34	ACTA2
KRT5	ESR1	PECAM1	COL1A1
KRT6	ESR2	ADGRL4	COL4A1
KRT14			
KRT17			

Table 4.5: Canonical Markers for Non-Immune Cell Type Assignment

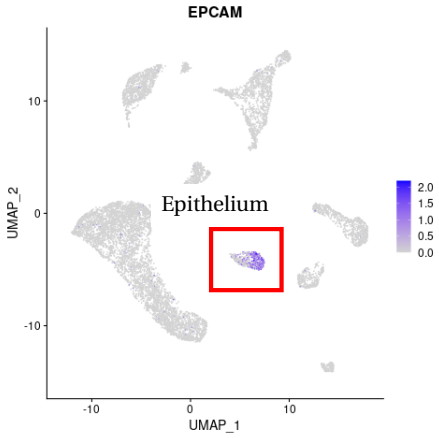
Cytotoxic T cells	Helper T cells	Plasma Cells	non-Plasma B cells
CD2	CD2	JCHAIN	CD19
CD3D	CD3D	IGHA1	
	CD25		

Table 4.6: Canonical Markers for Adaptive Immune Cell Type Assignment

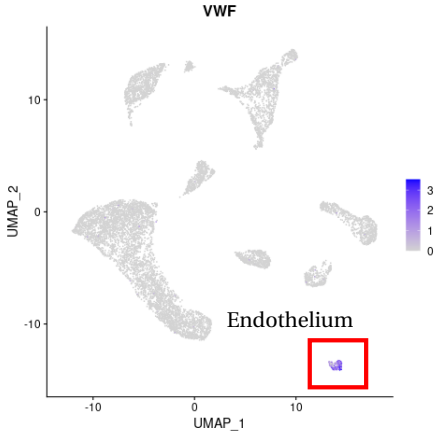
pDC	NK cells	Mast Cells	Myeloid Cells
IL3RA	NCAM1	KIT	ITGAM
CLEC4C	FCGR3A	TPSAB1	ITGAX
NRP1		TPSD1	CD14
			CD68
			FCGR3A

Table 4.7: Canonical Markers for Innate Immune Cell Type Assignment

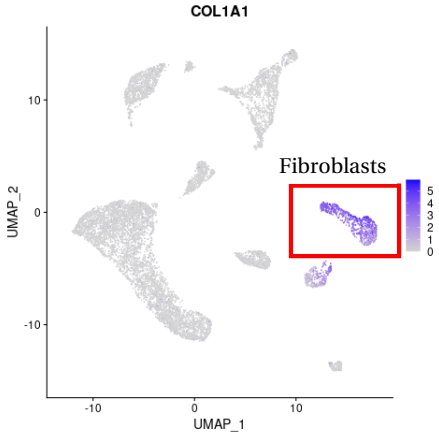
Cell type annotation using canonical marker expression for sample BC18 is shown in Figure 4.7.



(a)

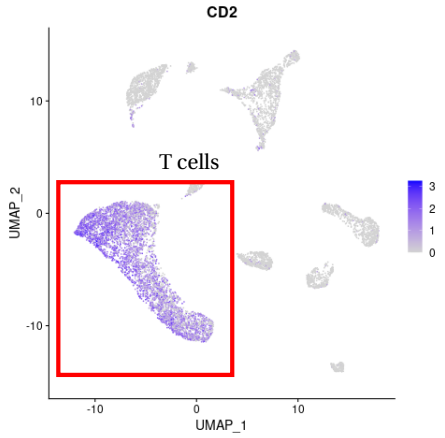


(b)

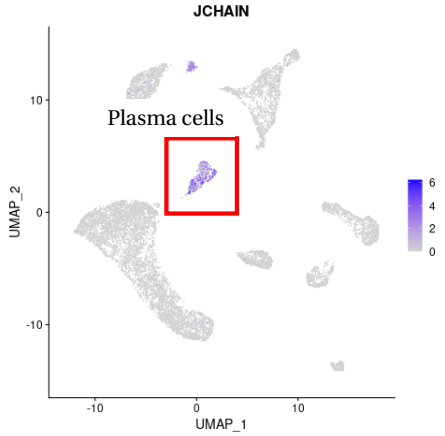


(c)

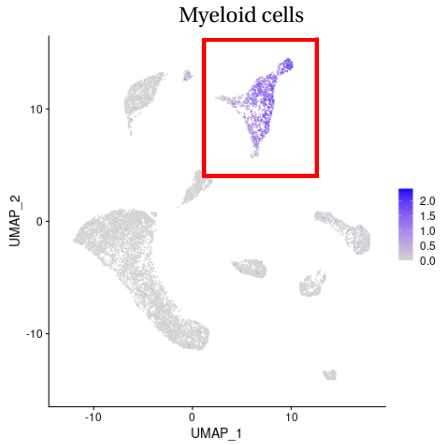
Figure 4.7: Manual Single Cell Annotation. Cell types are assigned to each cluster by expression of canonical markers. The canonical marker adopted for each cell type is displayed in the plot title. The legend details the expression level of the marker.
(a) Epithelial cells, (b) endothelial cells and (c) fibroblasts present in sample BC18.



(d)



(e)



(f)

Figure 4.7: Manual Single Cell Annotation. (d) T cells, (e) plasma cells and (f) myeloid cells present in sample BC18.

4.3.3.1 Integration of Unimodal Single Cell Datasets

The next step involves integration of separate patient samples to construct a single breast cancer atlas.

Harmony is a recently published algorithm enabling the integration of single cell datasets across different experimental batches and technology modalities [37]. The algorithm uses entropic-regularised soft clustering to maximise cluster diversity [37]. Batch effect is removed by applying a correction factor to each cluster under the assumption of linearity between the batch and response variables [37].

Optimal cluster assignment is calculated by iteratively applying the Expectation-Maximisation (EM) algorithm [86]. The EM algorithm is applied across many areas of statistical genomics. Results of sample integration using the Harmony algorithm are shown in Figure 4.8.

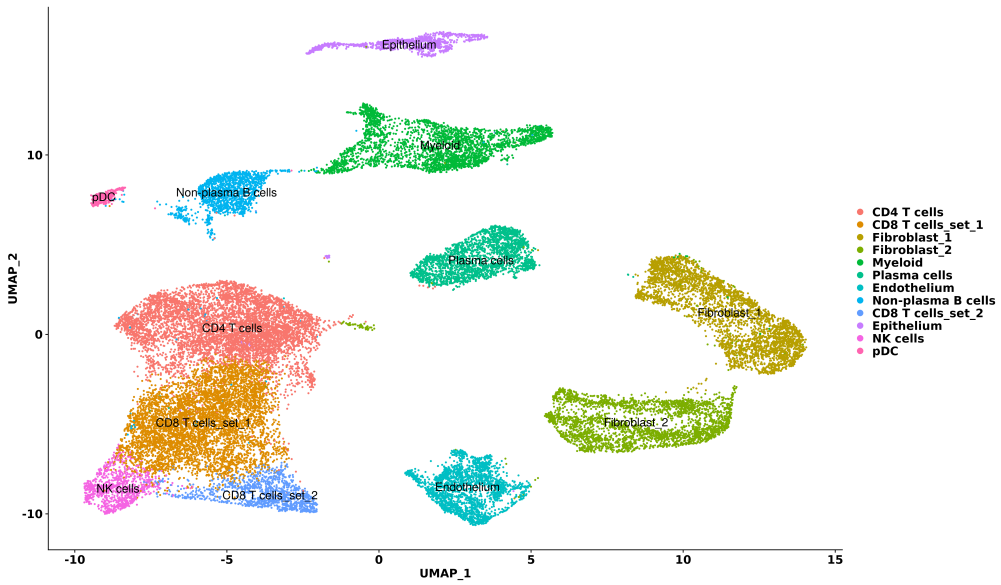


Figure 4.8: Single Cell Data Integration with Harmony. Integration of single cell data from three breast cancer samples. The axes denote the (x,y) coordinate position for each cell within the UMAP latent embedding space. Cell type heterogeneity is present across the three samples.

Figure 4.8 demonstrates the cell type heterogeneity of clinical breast cancer samples. The most dominant cell populations are T cells and fibroblasts. T cells exhibit sub-clusters which correspond to $CD8^+$ and $CD4^+$ T cells. Investigation of T cell heterogeneity may warrant future exploration.

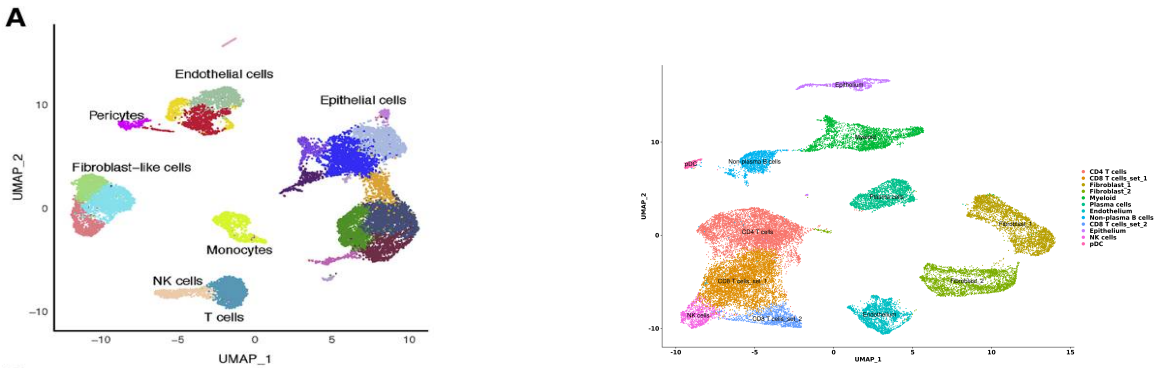
Heterogeneity between the fibroblast clusters may also warrant follow-up investigation. I observed changes in the surrounding breast tissue at the time of surgical excision for the collected patient samples. There was a marked desmoplastic response. Differential expression analysis between the fibroblast clusters can be performed followed by pathway analysis of the differentially expressed genes to identity potential functional consequences of the fibroblastic heterogeneity.

The tumour epithelial component is less extensive than expected. The tissue samples were collected intra-operatively directly from tumour tissue. Therefore, the epithelial component was expected to be the dominant cell population. It is possible the epithelial component was preferentially diminished due to technical factors related to the dissociation or single cell encapsulation process. Exploration of factors in the experimental protocol which may have resulted in selective epithelial cell loss may warrant further investigation.

Given the low sample numbers, the generalisability of the results is difficult to fully ascertain. The results appear to be consistent with the published literature [64].

The generated results were compared to scRNA-seq findings generated by Bhat-Nakshatri et al from normal human mammary tissue obtained as part of reduction mammoplasties. The UMAP embedding from normal mammary tissue is shown in Figure 4.9.

Comparison of the malignant vs normal mammary UMAP embeddings reveals several interesting observations. TNBC tissue appears to have a reduced epithelial and endothelial population alongside a marked expansion in the



(a) **(b)**

Figure 4.9: Comparison of the UMAP Embeddings of Normal and Malignant Mammary Tissue. Embedding of (a) normal mammary tissue (reproduced from [87]) and (b) TNBC tissue.

immune and fibroblast compartments compared to normal mammary tissue. In particular, the adaptive immune response is expanded in the malignant setting with heterogenous T and B cell subclusters. It suggests that the host immune response is subverted in breast cancer. However based on the presented results, it is unclear if it is a causative or reactive relationship. The temporal sequence of factors governing compositional transition from the normal to malignant state is an area of active investigation and leans towards a more multisystem perspective of both localised and metastatic breast cancer.

4.3.4 Future Considerations

4.3.4.1 Limitations of UMAP Plots

A key component in the interpretation of scRNA-seq data is the assumption that the PCA pre-conditioned UMAP embeddings offer meaningful and faithful latent representations of the data. There is emerging evidence that the theoretical propositions discussed in Subsection 4.3.2.3 do not hold in the real-world setting.

It is an important consideration for the single cell community and merits due discussion. I therefore present a critical appraisal of the pre-print by Chari et al [88].

Two components will be explored:

i. Data distortions generated by current popular embedding methods.

ii. Alternative semi-supervised structure-preserving latent representations.

During the discussion, I will focus on UMAP embeddings. The same general principles apply for t-SNE embeddings.

4.3.4.1.1 Data Distortions PCA-preconditioned UMAP embeddings form the backbone of current scRNA-seq analysis and interpretation [89]. Visualised closeness in the 2D embeddings is used to infer biological relationships and validate suggested transcriptional similarity between adjacent clusters [90]. However, these assumptions do not hold in practice. Furthermore, they ignore the effect of PCA-coupling on the robustness of the embedded representation [91].

The most important theoretical guidance to consider in this context is the Johnson-Lindenstrauss Lemma [92]. The Johnson-Lindenstrauss Lemma provides the sufficiency condition for low-distortion dimensionality reduction in Euclidean space [92]. It states that:

"Pairwise distances of m points can be preserved within a factor $(1 \pm \epsilon)$ for order $\frac{\log(m)}{\epsilon^2}$ dimensions" [92].

As an example, for a dataset containing 10,000 observations, distortion of pairwise distances within 20% accuracy can be achieved for a minimum of 1842 dimensions [88]. 10,000 cells is a common sample size seen in a standard single cell experiment. Therefore, the minimum number of dimensions required to preserve pairwise distances as guided by the Johnson-Lindenstrauss Lemma is much greater than the 2D UMAP embedding space.

Chiari et al proceed to quantify embedding induced distortions using real-world scRNA-seq data. A Seurat-integrated ex- and in-utero mouse embryo dataset was evaluated [93]. Pairwise distances between cells were calculated and then divided into pairwise distances of:

i. Low absolute value ('near and equidistant' set).

ii. High absolute value ('far and equidistant' set).

The mouse embryo dataset was selected because the 2D embedding of the ex-utero data was used to validate its representative capacity as a model for in-utero embryogenesis [88]. The authors restrict investigation to the chondrocyte and osteoblast cell populations [88].

Distortion of datapoints in the 'near and equidistant' and 'far and equidistant' groups is identified. Although the two sets of points exhibit distinct properties in the original high-dimensional space, they appear similarly clustered in the 2D embedding as demonstrated in Figure 4.10.

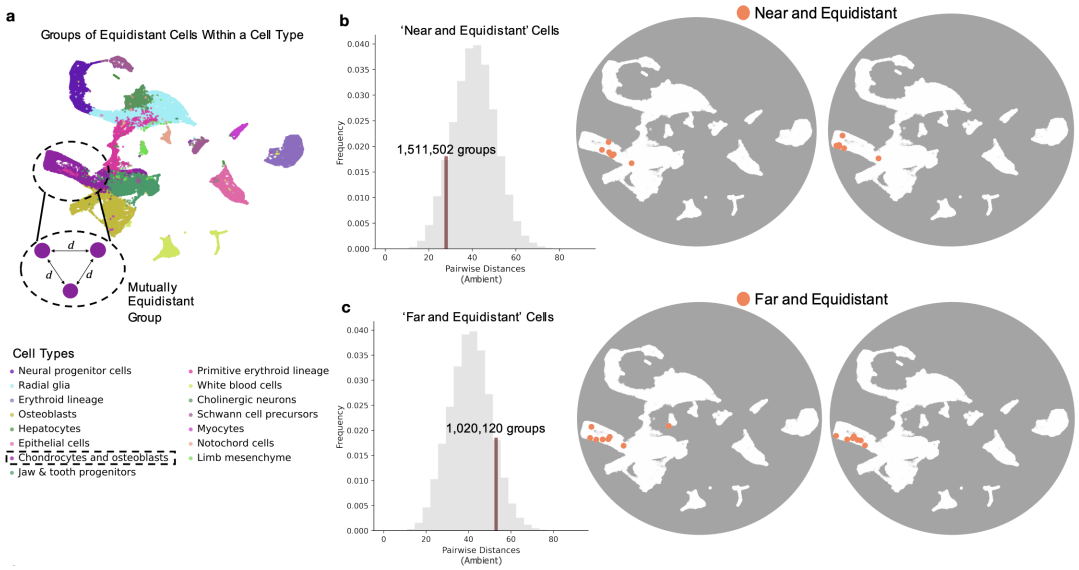


Figure 4.10: 2D UMAP Embeddings can introduce Data Distortions. Reproduced from [88].

As a control experiment, the authors engineer an autoencoder framework which preserves cell-to-cell distances from the original data space whilst fitting cells to an arbitrary user-defined shape [88]. The method is named 'Picasso' as homage to Pablo Picasso's skill in imitating artistic works [88].

The arbitrary shapes selected are:

i. Outline of the world map.

ii. The four-parameter von Neumann elephant.

The Picasso embeddings fitted to the ex-utero data are demonstrated in Figure 4.11.

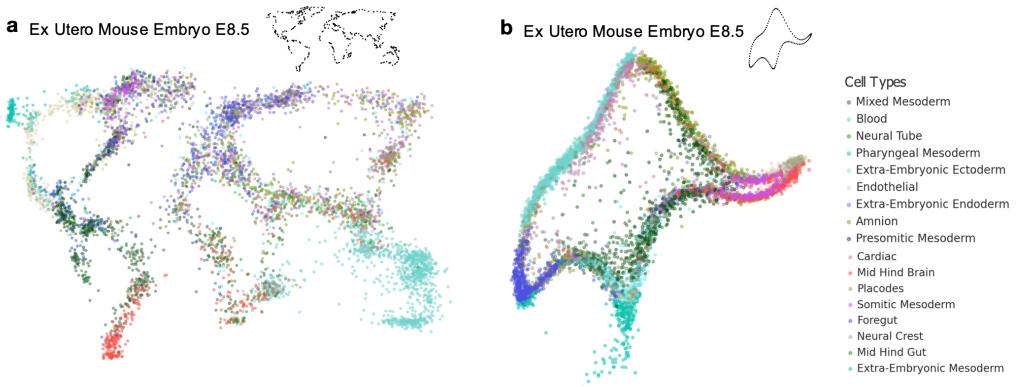


Figure 4.11: Picasso Embeddings of Ex Utero Mouse Embryo scRNA-seq Data. Embedding fitted to the shape of (a) the world map and (b) the von Neumann elephant. Reproduced from [88].

Correlation benchmarks for the Picasso embeddings are compared with baseline 2D embeddings as shown in Figure 4.12.

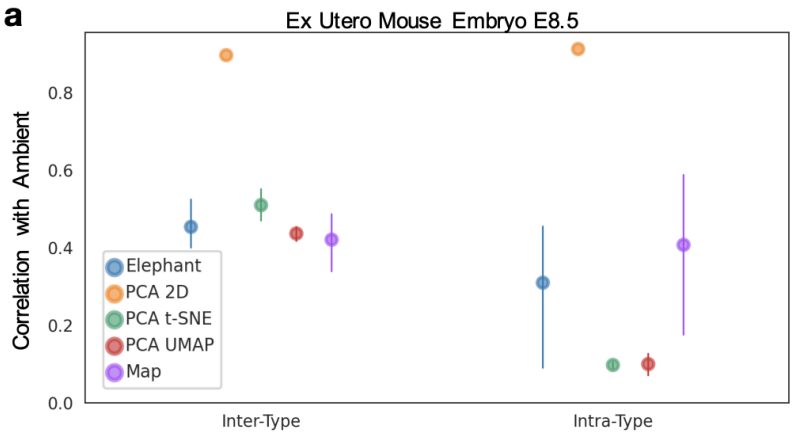


Figure 4.12: Correlation Benchmarks for Ex Utero Mouse Embryo scRNA-seq Data. Correlation metrics comparing Picasso embedding with popular embedding approaches. Reproduced from [88].

The correlation between inter-cluster points is comparable between the Picasso and UMAP embeddings [88]. In contrast, the correlation for intra-cluster points with Picasso outperformed the UMAP embedding [88]. Therefore, arbitrary user-defined shapes can represent inferred relationships with equal, or in some cases better, performance compared to 2D UMAP embeddings [88]. It suggests UMAP visualisations are not canonical. They should be interpreted

with caution [88].

4.3.4.1.2 Semi-Supervised Latent Representations Current methods for latent embedding are unsupervised. Based on the distortion findings discussed above, Chari et al propose a new semi-supervised autoencoder architecture.

It enables ground-truth biological relationships to guide latent learning by engineering multi-class multi-label data as input. Each cell can be assigned to greater than 2 classes for a range of key experimental factors, such as cell type, experimental condition, perturbation [88].

A two-layer autoencoder framework is adopted with a linear-decoder layer to facilitate interpretability of the latent class assignment [88]. The model is trained using the Adam optimisation algorithm on a combined reconstruction and label-based cost [88]. Standard methods of batch normalisation, ReLU activation and dropout regularisation are applied between the fully connected layers to facilitate training. The model architecture, called MCML, is shown in Figure 4.13.

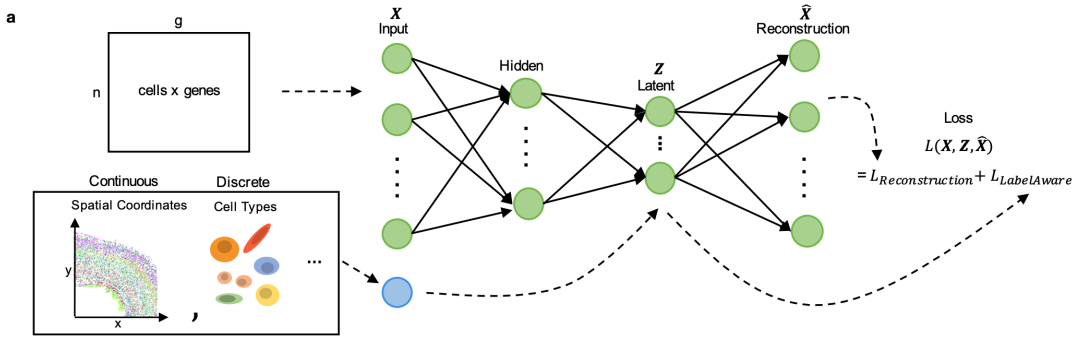


Figure 4.13: MCL Autoencoder Architecture. MCL consists of a two-layer autoencoder architecture with a combined reconstruction and label-aware cost function. Reproduced from [88].

MCML aims to cluster cells within the latent embedding in a manner which respects their biological origins, eg experimental condition, thereby improving the interpretability of inferred biological relationships.

When applied to the ex and in-utero mouse dataset, the MCML algorithm identified myocytes and hepatocytes as having the greatest distances between the in-utero and ex-utero batches within the latent embedding. It suggests ex-utero models of myocytes and hepatocytes are least representative of their respective in-utero biology. The finding was further supported by discordance in marker gene expression for these cell types between the ex and in-utero batches. The differentially expressed genes between the batches for myocytes and hepatocytes are shown in Figure 4.14.

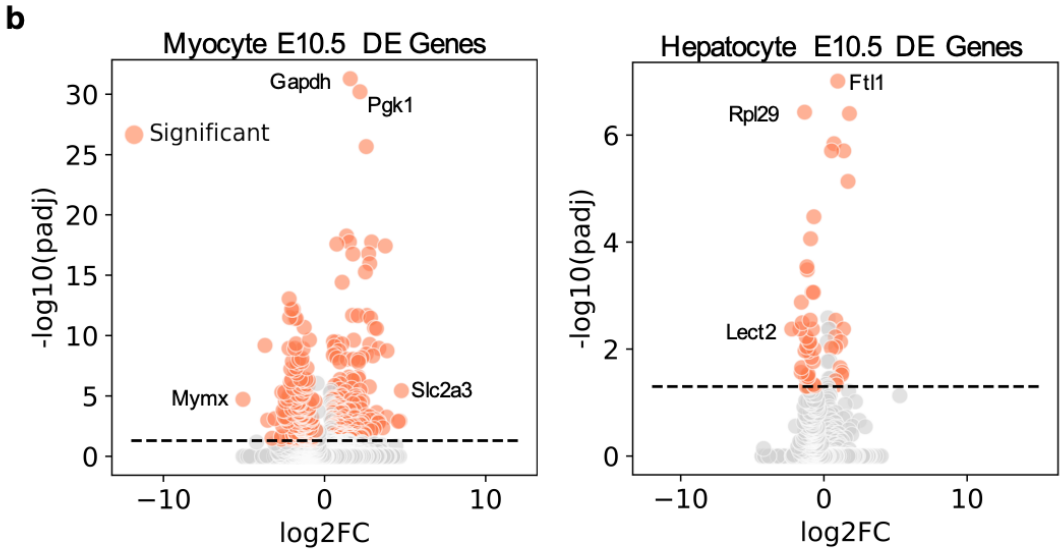


Figure 4.14: Differentially Expressed Genes. Myocytes and hepatocytes generated in ex-utero and in-utero models express distinct gene expression programs. Reproduced from [88].

In summary, there are limitations with currently adopted 2D UMAP visualisations. There are emerging alternative non-linear dimensionality reduction approaches which may more faithfully retain biologically relevant relationships.

4.4 Discussion

Breast cancer is a heterogenous disease [81]. The breast cancer ecoystem is characterised by communication between heterogeneous tumour cells, cells of innate and adaptive immunity, fibroblasts and endothelial cells [83]. The homotypic and heterotypic cellular relationships ultimately shape the tumour ecosystem and define disease progression and clinical outcomes [94]. Therefore, a comprehensive understanding of the tumour ecosystem is critical in identifying and developing new therapeutic approaches.

In the presented body of work, it has been demonstrated that scRNA-seq data can be successfully collected from a cohort of three patients in collaboration with a diverse research and clinical team. All tissue was collected prior to the administration of systemic chemotherapy or radiotherapy.

The results demonstrate the presence of cellular heterogeneity in breast cancer. Breast cancer is composed of a diversity of cell types: tumour epithelial cells, T and B cells, plasma cells, macrophages, fibroblasts, endothelial cells, NK cells and dendritic cells.

The results show that droplet based scRNA-seq can be applied to clinical samples. Following completion of data preprocessing, a pilot dataset has been generated which may be amenable to future exploration. The results from several patient samples can be aggregated in a batch-corrected manner using emerging algorithmic approaches.

The wider question on the use and interpretation of 2D UMAP plots in single cell analysis remains unanswered and requires future exploration.

4.4.1 Limitations

There are several limitations with the work. The low patient number (n=3) impacts on the generalisability of results. It is therefore not possible to explore nuanced aspects of biology with the current dataset.

The low tumor cell number was surprising. It is suggestive of potential technical bias in the experimental protocol. Future work may include identifying and optimising potential sources of bias in the sample processing pipeline. It is unclear if the factors resulting in epithelial deselection may also indirectly introduce bias on results from the non-epithelial compartments.

4.4.2 Future Directions

Future directions for investigation would include expanding the number of collected TNBC samples processed by scRNA-seq. A larger patient cohort would offer a more diverse and encompassing perspective on TNBC biology. It may improve the generalisability of results for future therapeutic development.

Breast cancer can be stratified according to the IntClust classification system. An expanded single cell atlas stratified according to this classification system would be invaluable to understanding the driving factors for heterogeneous treatment response and survival outcomes.

A further valuable study would be a comparison of pre- and post-chemotherapy breast cancer tissue at a single cell level. The study may provide detailed insight into the driving factors of treatment resistance, thereby offering a data-driven framework for developing new therapies.

In conclusion, one preliminary contribution has been made to the wider effort of deconvolving and understanding the breast cancer ecosystem at single cell resolution.

5

Multimodal Profiling of Breast Cancer

Contents

5.1 Multimodal Integration 124
5.1.1 Definition 124
5.1.2 Benefit 127
5.1.3 Challenges 128
5.2 Current Methods 129
5.2.1 Weighted Nearest Neighbour 129
5.2.2 MultiMap 134
5.2.3 Multi-Omics Factor Analysis v2 139
5.3 Future Methods 145
5.3.1 Multimodal Variational Autoencoders 146
5.3.2 Diffusion Models 148
5.4 Conclusion 159

5.1 Multimodal Integration

5.1.1 Definition

Single cell multimodal integration is the simultaneous measurement of data modalities across '-omic' space in same or paired samples. The potential span of integrated modalities is diverse, ranging from DNA, chromatin, RNA, protein and spatial location.

There are two classification systems for multimodal integration: practical and directional.

5.1.1.1 The Practical Classification System for Multimodal Integration

In some circumstances, multiple modalities are measured from the same cell at the same point in time using specific multimodal techniques, such as CITE-seq which simultaneously measures RNA and cell surface protein expression [95]. This is referred to as experimental multimodal integration.

Alternatively, multimodal integration can be performed using computational tools since the majority of single cell omics techniques are unimodal and tissue destructive. This is referred to as computational multimodal integration. In the latter case, different modalities are experimentally measured along serial planes from the same sample and then later computationally integrated.

There are benefits and disadvantages with each approach. Multimodal integration in experimental space offers the gold standard approach: paired measurements are taken from same cell at the same time point. However, the type of modalities available for integration are constrained by experimental availability, the techniques are bespoke requiring specialised experience and are expensive.

In contrast, multimodal integration in computational space is practically more achievable and so available across a wider range of research contexts.

Tissue destructive, single modality measurements are taken close together from within the same sample. However, intra-sample heterogeneity may make alignment in computational space difficult and introduce spurious alignments which are difficult to identify and control.

I will now present a critical appraisal of multimodal techniques which were anticipated to be explored in the absence of pandemic-related disrutpion.

5.1.1.2 The Directional Classification System for Multimodal Integration

Multimodal integration can be classified according to the direction in which differing modalities are aligned. There are three types as shown in Figure 5.1:

i. Horizontal integration

ii. Vertical integration

iii. Diagonal integration

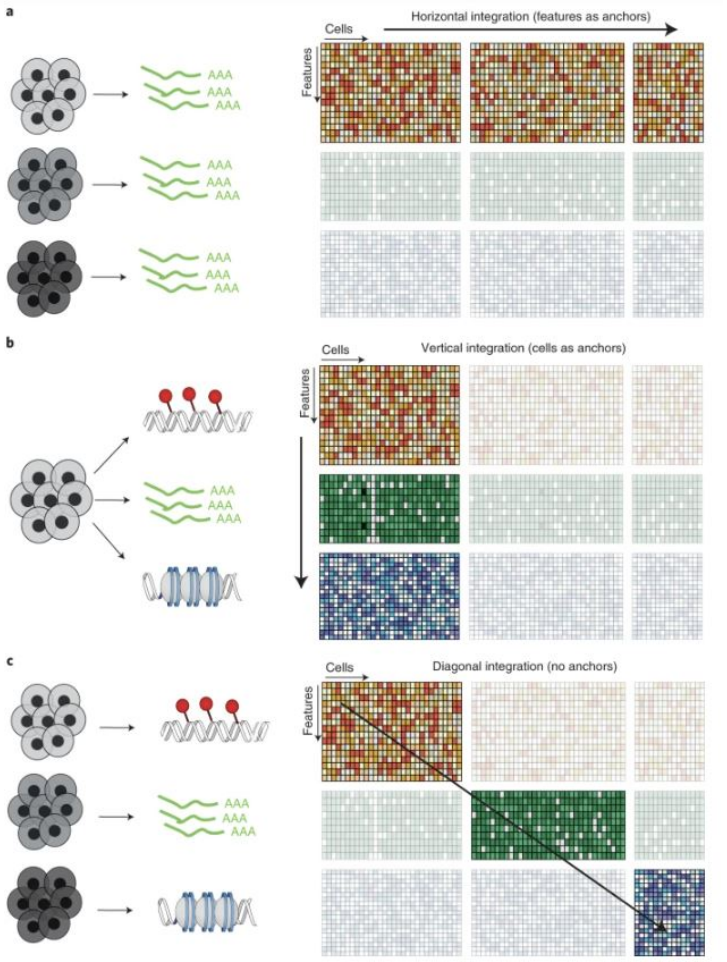


Figure 5.1: The directional classification system for multimodal integration. Reproduced from [96].

In horizontal integration, cells are aligned according to genomic anchors [96]. As an example, a set of common genes are shared between different samples measured using the same modality. In a certain light, this may be considered a type of batch correction. It will not be explored in more detail as an example of multimodal integration.

In vertical integration, cells serve as the alignment axis [96]. Multimodal measurements are taken from the same cell using tailored techniques, eg CITE-seq.

Finally, in diagonal integration, neither cells nor measurement modality are common between samples [96]. It represents the most commonly encountered example of multimodal integration but it is the most difficult to achieve and

to benchmark.

5.1.2 Benefit

Cellular state is a core concept underpinning biology. Characterisation of the cellular state and the identity of each cell within a multicellular hierarchial system will provide the foundational building blocks to better the understand the homeostatic and pathogenic processes of nature. It will offer insight into the dynamic and complex communication networks which regulate and drive tissue development, the maintenance of normal tissue equilibrium and the perturbations which account for and propagate disease.

Yet for such a foundational property, cellular state remains poorly defined. To date, there does not exist a systematic generalisable framework which captures all the explicit and implicit properties encompassed by such a foundational concept. The distinction between cellular state and cellular identity is unclear. The set of all potential cellular states and how they relate to each other is not certain. The factors guiding the transitions between states is unknown.

The insights needed to better understand cellular state and identity cannot be gleamed from a single modality. RNA alone or protein alone is insufficient. Instead, we must integrate across orthogonal biological axes, traversing DNA sequence, chromatin configuration, gene expression, protein level and spatial location in order to capture the requisite richness of information flow encompassed under the umbrella of cellular state. Therefore, the question of defining cellular state in essence becomes a need to establish interpretable multimodal integration and translation across biological domains.

The anticipated benefits would enable the field to:

i. Formally define the properties of cellular state

ii. Establish an universal topological map of cellular states

iii. Identify new cellular states

iv. Identify new and rare cell types

v. Better understand the link between shifts in cellular state and the development and progression of disease

We must integrate across domains in a biologically informed way. Experimental design should be guided by the anticipated axes of greatest information content and tailored to the context in which it is being applied. Technological constraints and availability should be a secondary consideration.

5.1.3 Challenges

There are several challenges to multimodal integration [97]. The challenges are:

i. Feature discrepancy between modalities

ii. Missing data

iii. Noisy data

iv. Difficulty in identifying the ground truth

iv. Data scale and the computational cost of multimodal integration

The above challenges are a reflection that data from different modalities come from different underlying statistical distributions and have different properties with different levels of noise and distinct batch-specific properties [98]. For example, gene expression will lie in a feature space fixed per species. The dimensionality of the gene expression vector space for humans is ~20,000. In contrast, the space for protein expression as measured by CITE-seq is smaller, typically of dimensionality ~200 and exhibits inter-experiment variability.

There are also practical considerations. The size of multimodal datasets are typically much larger than in the unimodal context. Managing large datasets requires tailored data management tools, access to a managed high performance computing infrastructure and additional software engineering skills.

5.2 Current Methods

I will present a summary of the following commonly used multimodal tools:

i. Weighted-Nearest Neighbor (WNN)

ii. MultiMap

iii. Multi-Omics Factor Analysis v2 (MOFA+)

5.2.1 Weighted Nearest Neighbour

Hao et al introduce a method of WNN analysis to learn a multimodal representation of cell identity. The tool involves four core steps:

a. Generation of an independent KNN graph for each modality

b. Intra- and inter-modality prediction

c. Determination of modality weights for each cell

d. Construction of a WNN graph

WNN analysis was applied to a CITE-seq dataset of 211,000 PBMCs integrating single cell gene and protein expression data.

5.2.1.1 Generation of a KNN Graph per Modality

Data from each modality was pre-processed individually using standard pre-processing approaches. scRNA-seq data was pre-processed in Seurat by normalisation, feature selection and dimensionality reduction by PCA. Single cell protein expression data was pre-processed by normalisation and PCA dimensionality reduction.

The pre-processing output was used to construct a KNN-graph. k, a hyper-parameter for KNN clustering, is the number of nearest neighbours. For most analysis, a default value of $k = 20$ was applied.

5.2.1.2 Intra- and Inter-Modality Prediction

Conceptually, the RNA and protein profile of a cell is predicted based on its surrounding neighbours using the predefined value of *k*. It is conducted in an intra-modality manner, ie the RNA profile of a cell is predicted based on the RNA profile of its neighbours, and in a inter-modality manner, ie the RNA profile of a cell is predicted based on the protein profile of its neighbours.

Intra-modality prediction for RNA expression is calculated by:

$$\hat{r}_{i,knn_r} = \frac{\sum_{j=1}^{k} r_{knn_{r,i,j}}}{k}$$

Inter-modality prediction for RNA expression is calculated by:

$$\hat{r}_{i,knn_p} = \frac{\sum_{j=1}^{k} r_{knn_{p,i,j}}}{k}$$

where:

r_i = L2-normalised PCA-reduced vector of gene expression for cell *i*

p_i = L2-normalised PCA-reduced vector of protein expression for cell *i*

The same principles apply for prediction of protein expression.

5.2.1.3 Determination of Modality Weights

The following component is a key step in WNN. It involves:

i. Calculation of distance between the observed and predicted expression using the Euclidean metric

ii. Conversion of distance to affinities using a kernel method

iii. Calculation of normalised affinity ratios within and between modalities

Hao et al apply a kernel method to quantify the local connectivity of the graph.

A kernel is a generalised dot product. For example, if we have two vectors $\mathbf{x}$ and $\mathbf{y}$ and a mapping $\varphi : \mathbb{R}_n \rightarrow \mathbb{R}_m$, then a kernel corresponds to $\mathbf{k}(x, y) =$

$\varphi(x)^T\varphi(y)$.

The kernel method is extremely useful because it allows one to perform calculations in an arbitrarily high-dimensional feature space, $\mathbb{R}_m$, without undergoing the computational cost of visiting this space. In $\mathbb{R}_m$, the vectors become linearly separable to allow classification.

Hao et al use an exponential kernel as utilised in the UMAP algorithm [77]. The exponent in this kernel consists of the maximum distance between a cell and its nearest neighbour set with an additional subtraction term used to provide numerical stability. Affinity ratios between modalities are calculated and softmax normalised, producing the cell-specific modality weights.

5.2.1.4 Construction of a WNN Graph

The resulting modality weight offers a new similarity distance between cells. The weights are used to construct the WNN based on a weighted combination of RNA and protein affinities.

An overview of the graph construction pipeline is shown in Figure 5.2.

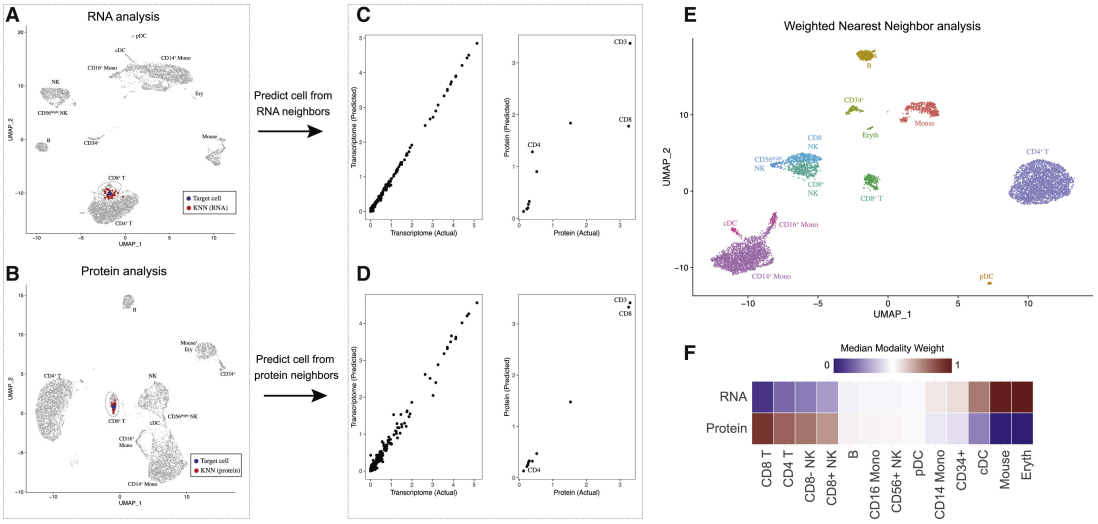


Figure 5.2: WNN Analysis. Reproduced from [99].

5.2.1.5 Application of WNN

WNN enabled characterisation of lymphoid heterogeneity amongst PMBCs. It was applied to time-series CITE-seq PBMC data from 8 patients enrolled on a HIV vaccine trial to construct a new multimodal atlas of human PBMCs.

Lymphoid cell states not normally reported from scRNA-seq data were identified. Specifically, sub-populations of double negative T cells, $\gamma\delta$ T cells and $CD8^+$ memory T cell populations, normally only reported in solid tissue, were identified.

It suggests multimodal data can give insight into a greater resolution and diversity of cellular states compared to unimodal data. In this specific context, it captured the transition between circulating lymphoid cells to tissue resident T cells, a state which has hithero been difficult to capture.

5.2.1.6 Limitations of WNN

WNN offers a valuable contribution to the single cell field. It is a rigorously developed tool which can easily be integrated into existing single cell workflows. Hao et al have also developed online visualisation tools to enable the wider community to access and use WNN for in-house unpublished data.

Nevertheless, there are some limitations with the WNN tool. At its core, WNN assigns weightings to each modality based on the predicted information content and essentially uses the most weighted modality to define cell state. From a purest perspective, multimodal integration should ideally incorporate the most informative sub-components of each modality to provide a truly holistic assessment of cellular state.

It is a particularly pertinent consideration for when dimensionality imbalance is present between modalities. It is likely dimensionality imbalance will be frequently encountered. From an algorithmic perspective, KNN clustering

is sensitive to changes in dimension size. It would be expected that WNN is biased towards preferentially weighting the higher dimension data modality. Investigation into the impact of dimensionality imbalance may be warranted.

Cells for which conflicting states are identified through multimodal investigation have the potential to offer the greatest biological insight into how we understand cellular state. Such examples may challenge and reshape the very definition of cellular state. WNN uses a weighted linear combination of modalities for final graph construction. It is therefore unlikely to identify cases of state incongruence across modalities.

The intention is for WNN to be employed as input for standard single cell visualisation tools, such as tSNE and UMAP. The use of 2D visualisation tools, although good from the perspective of facilitating human interpretability, introduce data distortions with the risk of introducing artefactual biological conclusions. It is an aspect of single cell analysis discussed in further detail in Chapter 4. It may be constructive to explore alternative tools to facilitate data interpretation.

There are more technically oriented considerations of the WNN algorithm. The key step in the algorithm is the determination of modality weights. Currently, the exponential kernel is adopted. The theoretical justification for selecting this kernel, as inspired by the UMAP algorithm, is unclear.

In most contexts across disciplines, the Gaussian kernel is used because it has well-studied, universally stable properties. It would be helpful to further explore the factors guiding the selection of the exponential kernel, the theoretical properties which favour its application in the multimodal integration context and compare its performance with a range of commonly used kernels. It is possible different kernels may be best suited for specific modality pairings.

5.2.2 MultiMap

A distinct but complementary technique is MultiMap published by Jain et al [98]. It is a multimodal generalisation of the UMAP algorithm. The multimodal graph is constructed in manifold space, $\boldsymbol{M}$, from which a low-dimensional embedding is then estimated.

There are three core steps to the algorithm:

i. A fuzzy set representation of the observed high-dimensional data is projected onto manifold space. Intuitively, one can consider a fuzzy set as a mathematical object which generalises soft clusters

ii. A joint neighborhood graph, the MultiGraph, is constructed on the manifold.

iii. MultiGraph is projected to a low-dimensional embedding space.

5.2.2.1 Data Projection onto a Manifold

MultiMap is a multimodal generalisation of a manifold. It assumes that the observed data lies uniformly on an underlying manifold. Multimodality is accommodated by assuming that the underlying manifold contains multiple distinct ambient spaces.

However, it does not assume that data between modalities arise from the same feature space. It does not assume equal dimensionality between data from different modalities nor that distance can be defined between every pair of points.

5.2.2.2 Construction of MultiGraph

A key component in constructing the joint neighborhood graph involves determining distances between data points both within and between modalities. Intra-modality distance is calculated by normalising distance with respect to

a neighborhood distance specific to the dataset. Inter-modality distance is calculated by first mapping the data into a shared feature space, identifying the neighborhood parameter in this shared space and then normalising distance with respect to this neighborhood parameter.

The above calculated distances are then used to construct the neighborhood graph in $\boldsymbol{M}$, referred to as the MultiGraph.

5.2.2.3 Construction of Embedding Space

MultiGraph is then mapped to a low-dimensional embedding space by minimising the cross-entropy loss between the manifold and embedding space using stochastic gradient descent optimisation.

Downstream analysis and visualisation can then be performed on either the embedding or manifold space, offering user-specific flexibility.

A schematic of MultiMap is shown in Figure 5.3.

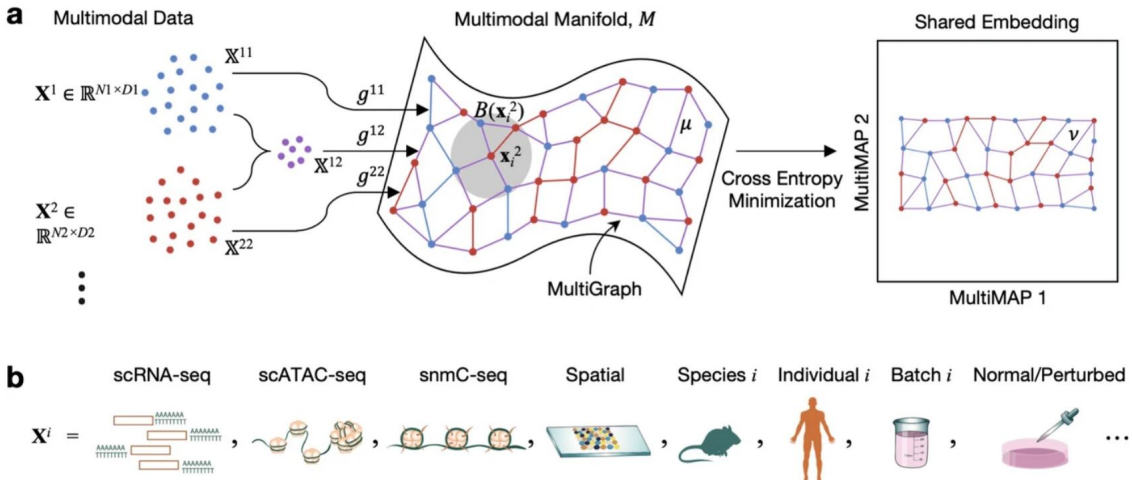


Figure 5.3: MultiMap Algorithm. Reproduced from [98].

5.2.2.4 Application

Jain et al apply MultiMap to a range of synthetic and experimentally-generated biological datasets. One of the applications included integrating Drop-seq scRNA-

seq and STARmap in situ spatial data of the mouse frontal cortex. I will focus on this application since it is most directly linked with the objectives of my DPhil.

The datasets demonstrate differences in feature size (Drop-seq encompasses the entire transcriptome and STARmap targest 1020 genes) and the number of cells (71640 for Dropseq and 2137 for STARmap). A summary of the results is shown in Figure 5.4.

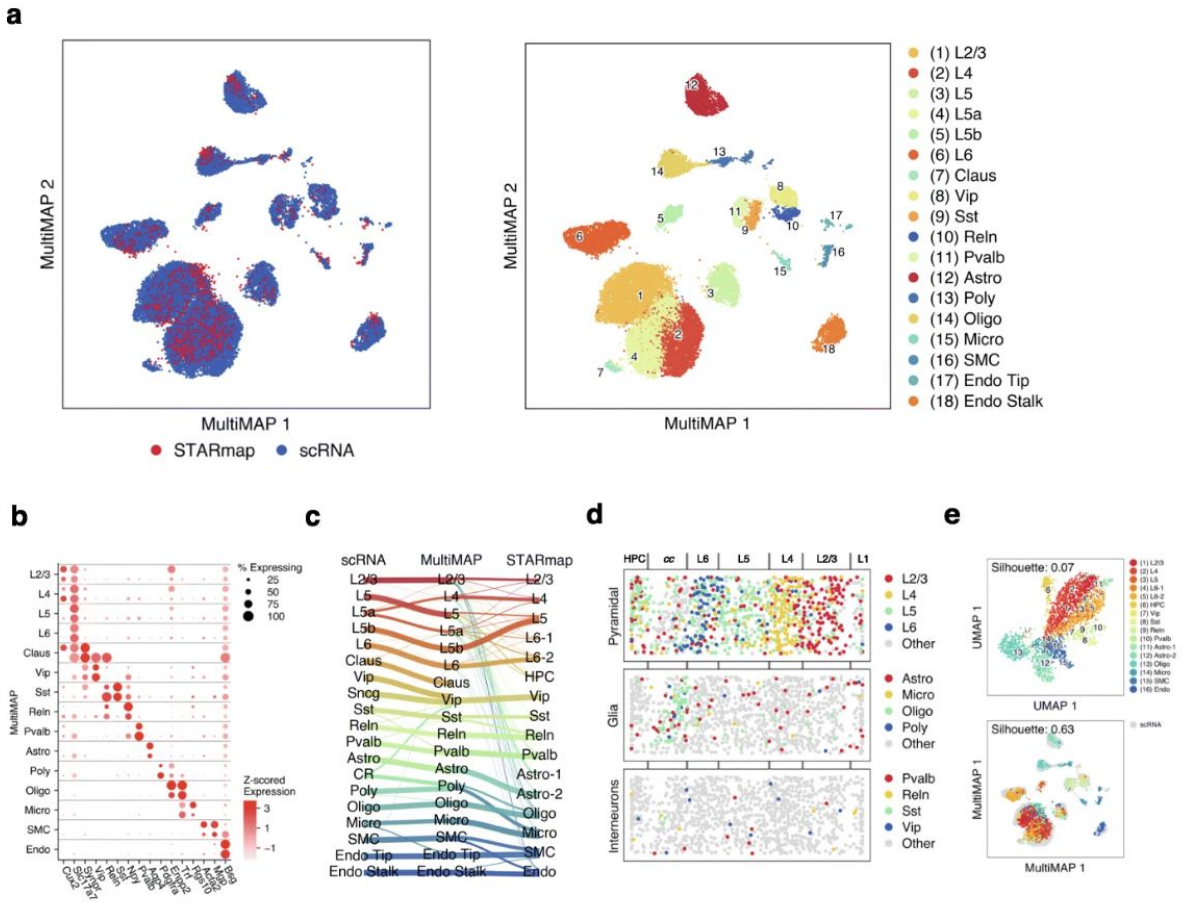


Figure 5.4: Multimap Integration of Spatial Transcriptomics and Gene Expression Data. Reproduced from [98].

Clustering using MultiGraph reproduced clusters corresponding to established cell types (Figures 5.4a and 5.4b).

MultiMap also enabled reannotation of a cell type. It was previously believed cortical layer 4 neurons do not exist in the murine frontal cortex [100]. The layer 4 neurons were identified on the STARmap data. Using multimodal integration, they were also later identified in scRNA-seq after being re-annotated from layer 5 to layer 4 neurons (Figure 5.4c). Therefore, MultiMap has demonstrable capacity to identify new cell types.

Multimap integration also improved clustering quality as determined by an increase in the Silhouette score of the embedding space (Figure 5.4e), and enabled spatial localisation of cell types in the STARmap data in a manner consistent with known cortical architecture (Figure 5.4d).

5.2.2.5 Advantages

There are several advantages to MultiMap. Unlike WNN in which data is PCA-reduced, MultiMap enables non-linear mappings. It can incorporate information from features across all modalities to improve the quality of data integration into the common embedding space. It is robust to working with datasets of varying dimensionality across modalities.

Furthermore, it is computationally scalable because the optimisation objective for mapping into the embedding space scales linearly with the number of data points, $\mathcal{O}(n)$. In contrast, methods such as WNN ran into out-of-memory errors despite access to 218 GB RAM [99].

5.2.2.6 Limitations

MultiMap is based on the Multimodal Manifold hypothesis. The hypothesis has two assumptions:

i. If multimodal data is considered to represent the same underlying system, then data from different modalities is assumed to be uniformly distributed on the shared manifold.

ii. The multimodal manifold is assumed to demonstrate coordinate invariance. If data points lie close together in a manifold space, then the same data points will lie close together in the empirical coordinate system.

Assumptions one and two are strong. The original UMAP publication by McInnes [77] explicitly discusses that real-world data will rarely be uniformly distributed on a manifold. The consequential algorithmic challenges and potential avenues to, at least partly, maintain local connectivity on an non-uniformly distributed data are explored in the original publication.

If challenges with data non-uniformity are present in the unimodal setting, it would intuitively be expected such limitations will be present or possibly amplified in the multimodal context. Therefore, it seems both appropriate and necessary to formally prove that real-world multimodal data can and does lie uniformly in a manifold space.

Assumption two is important. MultiMap estimates the geodesic distance between data points. It is a difficult problem because it relies on knowing the Riemannian metric which is unknown in the naive case. Therefore, Jain et al use assumption two to render calculation of geodesic distance into a tractable computation. Without assumption two, it would otherwise be extremely challenging to construct MultiGraph, a core component of the MultiMap algorithm.

The assumption of coordinate invariance may require further exploration. There is experimental evidence that coordinate invariance does not hold when mapping between high-dimensional and manifold space as discussed in chapter 4. The implications for MultiMap may warrant investigation.

Finally, there is the general question of applying Manifold theory in biology which could be considered in a partly philosophical light. A manifold is a specific

mathematical construct. It is a topological space that is locally Euclidean meaning that around every data point, there is a neighborhood which is topologically the same as a open unit ball in $\mathbb{R}_n$ [101].

The local Euclidean property is useful because it means a manifold can be a well-behaved type of space to work in. However, it also places a responsibility on every researcher applying Manifold theory to different disciplines. The first step of analysis should include a formal and mathematically rigorous demonstration that the data inherently retains this local Euclidean property. It is a not a data property which should be forced. If the property does not hold, then alternative analytical approaches should be pursued.

The UMAP algorithm and related techniques are becoming increasingly used across a range of single cell tools. It may be constructive to pause and reflect fully on the theoretical justification for becoming increasingly dependent on this branch of Mathematics. Every algorithm has its strengths and weaknesses. The applications best suited to manifold-based tools remain to be determined.

5.2.3 Multi-Omics Factor Analysis v2

MOFA+ involves the application of Bayesian statistical principles to matrix factorisation in the multimodal setting [102]. The input data consists of stacks as shown in Figure 5.5. Vertical stacks of matrices correspond to multimodal views in the same cell (View 1 - M) and horizontal stacks of matrices correspond to the same multimodal observations across different cell batches (Group 1 - G).

The assumption of matrix factorisation is that the observed data is a realisation of a lower-dimensional representation expressed by K latent factors. The latent factors capture the axes of variance in the global data. The latent factors are linked from the observed high-dimensional space to the underlying low-dimensional representation through a matrix of weights. The weight matrix

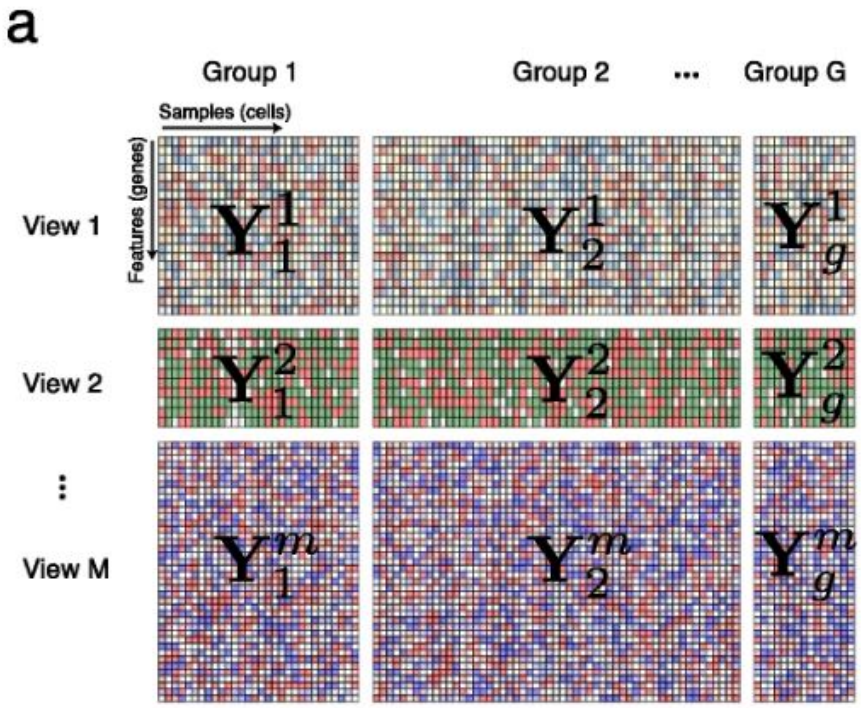


Figure 5.5: Structure of data input for MOFA+. Reproduced from [102].

connects each feature, such as a gene, to each latent factor and quantifies the strength of this connection [102].

Formally, matrix decomposition can be expressed as:

$$\boldsymbol{Y}_{gm} = \boldsymbol{Z}_g \boldsymbol{W}_m^T + \epsilon_{gm}$$

where $\boldsymbol{Y}_{gm}$ is the observation matrix, $\boldsymbol{W}_m$ is the weight matrix, $\boldsymbol{Z}_g$ is the factor matrix and ϵ_{gm} accounts for stochastic noise.

The defining property of the MOFA+ algorithm is the application of hierarchial priors to both the factor matrix, Z and the weight matrix, W [102]. A prior imposes a pre-defined statistical distribution on unknown model parameters. A hierarchial prior is a two-layer prior, ie a prior on a prior. It offers greater flexibility to engineer a better tailored model. It also acts as a form of model regularisation.

In MOFA+, the two layers of priors consist of:

i. Automatic Relevance Determination (ARD) prior to model factor activity across

both data modalities and sample batches [102]. It is a type of Bayesian ridge regression which enables input variables to be ranked based on their importance to predicting the output.

ii. Spike-and-slab prior to induce sparsity to the factor and weight matrix [102]. Sparsity helps with interpretability since it pushes the contribution of latent factors in most cells across most features to zero, thereby enabling the researcher to focus most of their efforts on the latent factors with a relevant connection to the inferred underlying generative process.

The graphical model for MOFA+ is shown in Figure 5.6.

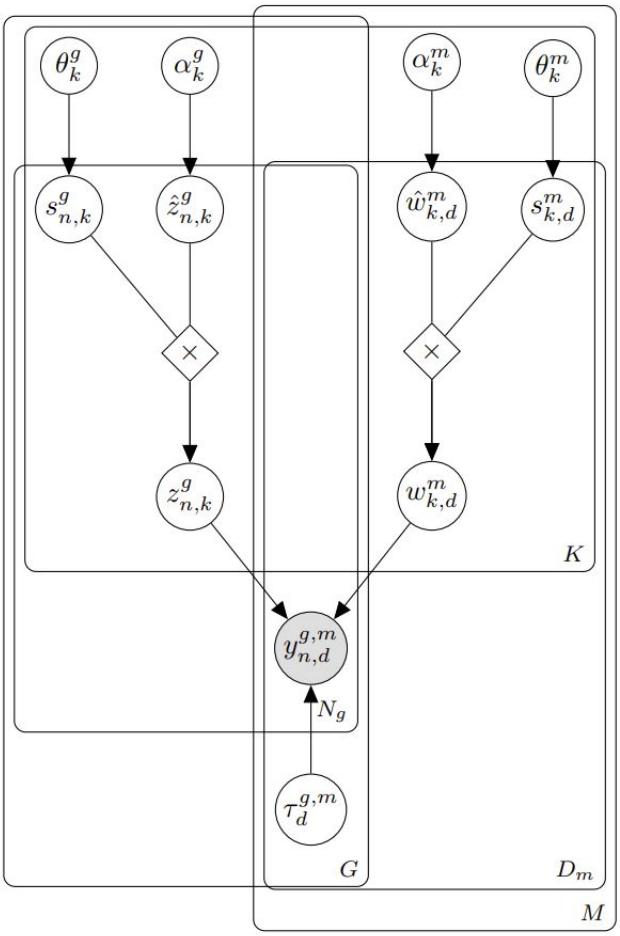


Figure 5.6: Graphical model for MOFA+ multimodal matrix factorisation. Reproduced from [102].

Grey circles are observed variable. White circles are latent variables. There are five plates, each representing a separate dimension. M are the number of modalities, D_m for the number of features in each modality, K for the number of factors, G for the number of groups and N_g for the number of samples in group g [102].

The graphical model shows that in essence, MOFA+ uses hierarchial priors to control how sparsity is introduced across modalities, features and datasets to then aid learning via Stochastic Variational Inference (SVI). The learning process is discussed in Section 5.2.3.1.

5.2.3.1 Model Training

The MOFA+ model is trained using SVI [102]. SVI is adopted because model inference in the probabilistic setting is often intractable. Therefore, the aim is to identify good approximate solutions.

Sampling based methods, such as Markov Chain Monte-Carlo (MCMC), are popular and are guaranteed to find a globally optimal solution. However, they have a key limitation. They are very slow and computationally expensive to train. Consequently, it may not be possible to identify a good solution in polynomial time.

Alternatively, the variational family of algorithms can be used for approximate inference. Variational inference recasts inference as an optimisation problem. The original distribution p is intractable. Instead, one solves over a class of tractable variational distributions Q. The aim is to identify a $q \in Q$ which best approximates p and then use q for downstream inference.

Unlike sampling-based methods, variational inference will never find the global optimal solution. However, it is guaranteed to converge with identifiable bounds on the accuracy of convergence, ie the Evidence Lower Bound (ELBO).

SVI is a type of VI. Rather than using the entire dataset for training, SVI uses smaller random batches of data to optimise the parameters of the variational distribution. Therefore, it is computationally more efficient and quicker to train.

The optimisation objective for MOFA+ is:

$$x^{(t+1)} = x^{(t)} + \rho^{(t)} \nabla F(x^{(t)})$$

where x are the variables to be inferred, *F(x)* is the ELBO and $\rho(t)$ is the step size which determines the 'speed' of model training [102]. $\rho(t)$ is an adaptive step size, adjusted at each training iteration [102]. MOFA+ is designed to enable GPU-accelerated training for working with large datasets.

5.2.3.2 Variance Decomposition

The above training process identifies the optimal decomposition of the observed data matrices into the underlying factor and weight matrices. To facilitate interpretability with the factorisation output, it is possible to identify the variance explained by each latent factor k across modalities and sample groups [102].

5.2.3.3 Application of MOFA+

MOFA+ was applied to investigate the transcriptional heterogeneity of mouse embryogenesis using time-series scRNA-seq data. Seven latent factors were identified which account for 35 - 55% of transcriptional cell-to-cell variance [102].

It was possible to attribute factors to specific cell states. Factors 1 and 2 represent extra-embryonic cell types and factor 5 represents epiblast to mesodermal transition via the primitive streak. For each factor, the top feature weights disproportionally represented lineage-specific markers [102].

Further insights of embryogenesis were inferred by MOFA+. Variance attributed to factor 1, a factor representing an extra-embryonic state, remained

constant throughout development [102]. It suggests that extra-embyronic fate commitment occurs early during embryogenesis.

MOFA+ was also applied to investigate the epigenetic profile of neurons by comparing methylation signatures in CpG and non-CpG islands across different brain regions [102]. Cells were grouped according to their position in the cortex: deep, middle and superficial cortical layers. Factor 1 distinguished between excitatory and inhibitory neurons. Factor 3 showed greater diversity amongst excitatory neurons situated in the deep cortical layer compared to superficial layer, thereby providing evidence for spatially-driven neuronal heterogeneity [102].

5.2.3.4 Limitations of MOFA+

There are several limitations to MOFA+. MOFA+ can be considered to be a generalisation of sparse PCA. Like PCA, MOFA+ can only identify linear directions of maximum variance, thereby imposing an assumption of linearly on the underlying model. Linearity can be useful since it aids with improving interpretability. However, it can be restrictive in biology where non-linear relationships and mappings are widespread.

As a statistical model, it models genes as independent variables. In practice, gene regulatory networks exhibit a high degree of correlation between features. Therefore, the assumption of independence may not be valid.

From a practical viewpoint, MOFA+ can only be applied when multimodal measurements are taken from the same cell. It has experimental implications since specific multimodal assays, such as CITE-seq, must be used in the data generating phase. These multimodal assays require specialist expertise, are best conducted in an environment with established multimodal expertise and are expensive.

5.3 Future Methods

Multimodal integration should have its roots in Representation Learning. The tools of Representation Learning are naturally synergistic and aligned with the modelling requirements for multimodal integration.

I propose a more foundational approach to the challenge of multimodal integration through a critical appraisal of key ML inspired multimodal papers. In particular, inspiration has been taken primarily from the field of deep generative modelling.

I will focus on two types of models:

1. Multimodal variational autoencoders (MMVAE)
2. Diffusion models

From a practical perspective, the development of techniques for multimodal integration will require two distinct but interlinked components:

i. Algorithms which enable multimodal integration.

ii. Approaches which transform the output from multimodal integration into an interpretable output. Much less progress has been made on this front, partly due to its technical and theoretical difficulties. However it is, at the very least, an equally important domain requiring advance. Multimodal integration which is difficult to translate into real-world translational significance will limit the potential for scientific discovery.

In the general multimodal context, text-image translation has been an area for which recent practice changing advances have been made. I propose these advances could guide multimodal integration in the biological domain. The core principles behind image-text translation are broadly similar to the requirements of multimodal translation in biology. Domain-specific adjustments tailored to the nuances of WGS, RNA-seq, ATAC-seq, spatial transcriptomics and proteomics data will undoubtedly be required.

5.3.1 Multimodal Variational Autoencoders

The aim of multimodal modeling is to learn generalisable representations of the key information content shared between different modalities which in turn captures semantically meaningful knowledge. The flow of information is bidirectional: from observation to representation and vice versa [103].

Variational autoencoders (VAE) provide the fundamental tools required for such an endeavour [104] but require optimisation and refinement for application in the multimodal context.

The key criteria required for a generative model in the multimodal context is [103]:

i. Latent factorisation: the latent space factorises into subspaces which contain vector spaces shared between modalities as well as modality specific vector spaces.

ii. Coherent joint generation: generation from the same latent vector exhibits coherence across modalities.

iii. Coherent cross generation: data generated for one modality by conditioning on a second modality is coherent.

iv. Synergy: the quality of the shared representation improves due to learning across modalities.

Every multimodal model should be benchmarked according to these four criteria. Assessment metrics can be associated with each criteria to facilitate a consistent and reproducible method for model evaluation.

In the work of Shi et al, a Mixure of experts Multimodal VAE (MMVAE) is proposed for multimodal integration [103]. A variational posterior consisting of a sum of Mixture of Experts (MoE) encompassing the broad range of modalities is trained to both jointly embed and generate multimodal observations.

Conceptually, the optimisation objective is the same as a standard VAE:

the ELBO is optimised using SGD. Modifications are made to the optimisation objective to provide a tighter lower bound on the loss by using an appropriately weighted multisample estimator (such a model is known as an Importance Weighted Autoencoder) which is extended across modalities through stratified sampling to ensure pan-modality training gradients are equal.

A Laplace prior-posterior pair is used to model the distribution of the observed and latent variables. The standard Gaussian-Gaussian conjugate pair is not used because the Laplace distribution offers comparable distributional flexibility with the added ability to break the rotational invariance of the Gaussian in order to facilitate latent factorisation [105].

To assess model performance, Shi et al explore model performance on the Caltech-UCSD Birds (CUB) dataset. It is a challenging image-language dataset since the images are detailed and the caption descriptions are succinct. To improve the computational burden of model training and sample generation, the dimensionality of the images and text are reduced by working within the feature space of a pretrained CNN and using word embeddings respectively.

With MMVAE, Shi et al demonstrate meaningful reconstructions can be generated across paired images and text (Figure 5.7):

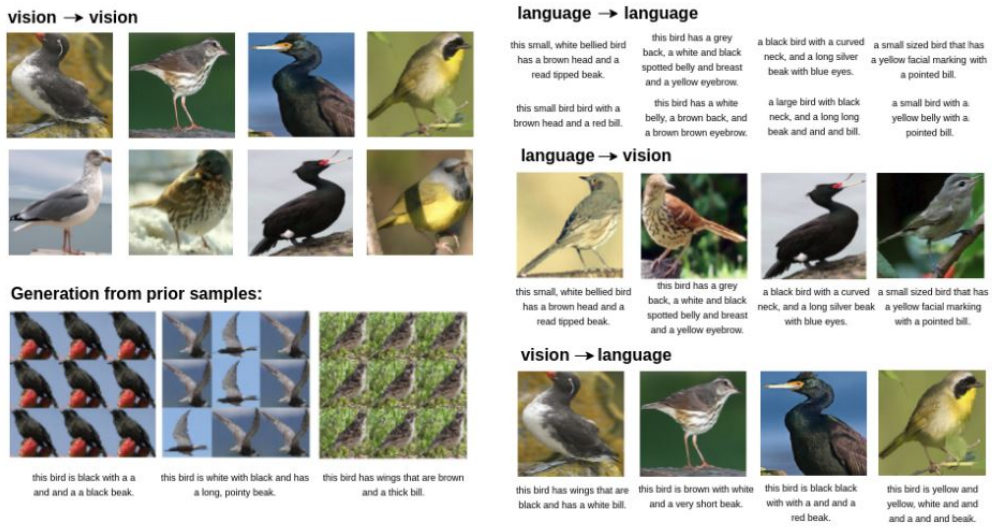


Figure 5.7: MMVAE generates meaningful cross-modality reconstructions. Reproduced from [103].

Figure 5.8 shows that the joint and cross generation performance is coherent. The correlation either from a common latent point or between modalities is consistently higher than the correlation performance for comparative models such as a Multimodal VAE (MVAE) [106]. However, the cross-modality correlation value still remains low, at a maximum of 0.135 and may benefit from further exploration.

Table 4: Correlation of Image (I)-Sentence (S) pair for joint and cross generation.

	Joint	Cross (I →S)	Cross (S →I)	Ground Truth
MMVAE	0.263	0.104	0.135	0.273
MVAE	−0.095	0.011	−0.013	

Figure 5.8: Correlation values for MMVAE. Reproduced from [103].

5.3.2 Diffusion Models

Diffusion models are the next generation of generative models which could have application in the multimodal context. They are probabilistic denoising models.

Diffusion models are conceptually similar to VAEs. Both types of models project data into a latent space and then use an optimisation objective to recover

the information content. However, VAEs suffer from challenges in modelling the ideal loss function. They can therefore be difficult to train and produce suboptimal reconstructions. At a high level, diffusion models address this limitation by modelling the noise distribution instead of the data distribution as in VAEs. Modelling the noise distribution turns out to be mathematically more stable.

Diffusion models have their origin in non-equilibrium thermodynamics, a branch of Statistical Physics which deals with entropic transport processes in systems which are evolving over time or space. The core tenant of diffusion models involves using a Markov chain to convert a well-characterised known distribution, eg a Gaussian distribution, into a complex target distribution.

It involves two steps: the forward process and the reverse process as shown in Figure 5.9.

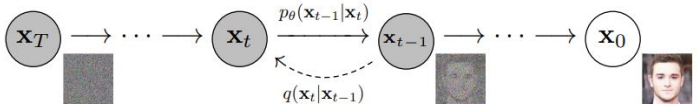


Figure 5.9: Directed Graphical Model for Forward and Reverse Diffusion Processes. Reproduced from [41].

In Figure 5.9, the forward process, $q(x_t|x_{t-1}$, is depicted by the dashed arrows. It involves taking the original data, a face in the example shown, and injecting noise with t steps to generate a noise distribution, x_T. The forward process does not involve training.

The reverse process, $p_\theta(x_{t-1}|x_t)$ involves taking the noise distribution to generate the original image as shown by the solid lines. The reverse process requires estimating the previous less noisy state given the current state which necessitates knowledge of the gradients for the previous steps. Therefore, unlike the forward process, the reverse process requires training a model.

The most commonly used model is the UNET [107], an encoder-decoder neural network which is widely used throughout Computer Vision. The UNET is

useful in this application because the input and output layers are of equal dimension and equal to the size of the data dimensions. The final layer produces two outputs, mean and variance, which is required for reconstructing a distribution.

The mathematical derivation of optimisation objective for the reverse process is elegant. It is easy to implement since it has been reformulated as a linear combination of KL-divergences together with an easier reparameterisation of the mean [108]. A detailed discussion of the derivation of the optimisation objective lies outside the scope of the thesis.

5.3.2.1 Application of Diffusion Models

As a proof of principle, I explored the multimodal generative capacity of diffusion models for application in the biomedical domain. I explored a range of commercially available and open-source options: DALL.E 2, Midjourney and Stable Diffusion. These models can conduct text-to-image translation.

A selection of five texts were identified focusing on histological descriptions of colorectal cancer or descriptions of spatial transcriptomics results. The texts are representative of descriptions commonly found in clinical pathology reports. Some of the texts, Text ID 4 and 5, are descriptions taken directly from the published scientific literature. The text descriptions are shown in Table 5.1.

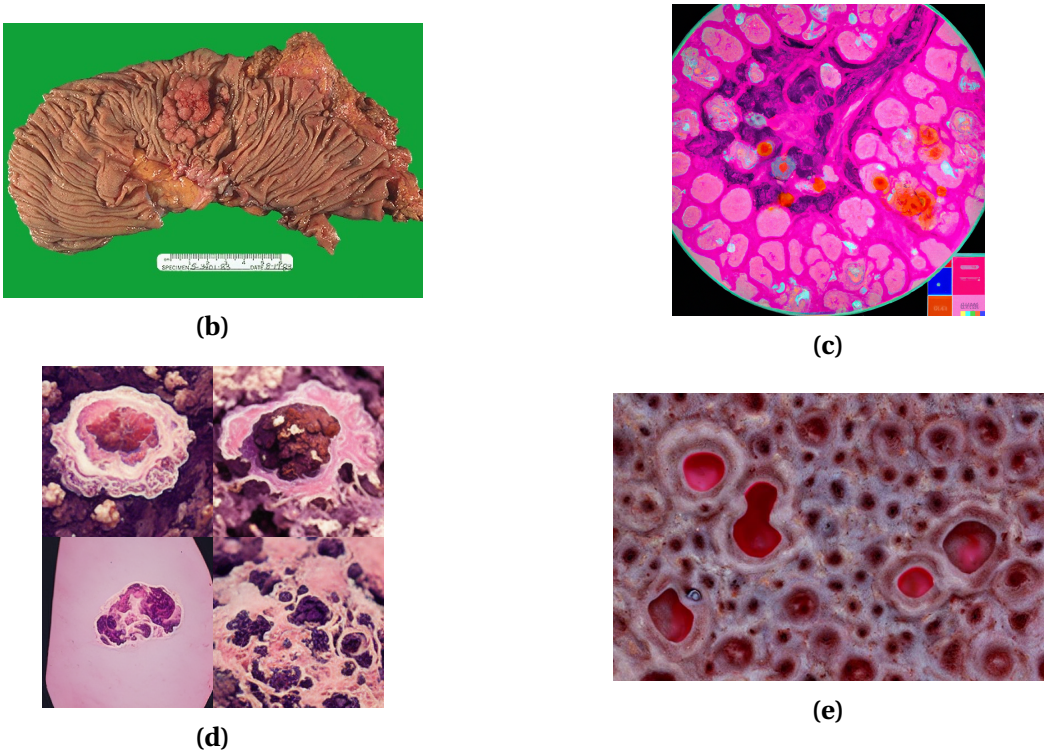
Text ID	Text	Source
1	This adenocarcinoma is arising in a villous adenoma. The surface of the neoplasm is polypoid and reddish pink. Hemorrhage from the surface of the tumor provides for detection with a stool test for occult blood. This neoplasm was located in the sigmoid colon, just out of reach of digital examination, but easily visualized with sigmoidoscopy.	[109]
2	The edge of the carcinoma arising in the villous adenoma is seen here. The neoplastic glands are long and frond-like, similar to those seen in a villous adenoma. The growth is primarily exophytic (outward into the lumen) and invasion is not seen at this point.	[109]
3	The tumour cells are polygonal displaying abundant cytoplasm, hyperchromatic enlarged nuclei with high N/C ratio, prominent nucleoli and nuclear irregularity/marked nuclear atypia. The tumour cells are forming nests, pseudo gland formation,trabecular and macro-trabecular patterns.	[110]
4	Tumor cells in nascent carcinomas (C-CIA) appear to be highly proliferative, clustering with Ki67 (cluster 1), while having weak spatial associations with immune cells (clusters 2), a trend that continues in CRC. Compared to benign adenomas, tumor cells in CRC have significantly weaker spatial associations with cytotoxic T cells, and significantly lower abundances of cytotoxic T cells and TNF-α, suggesting that the anti-tumor response has been suppressed in CRC, which frees tumor cells to divide at higher rates (increased CK/Ki67 spatial association).	[111]
5	Metastatic subclones occupied distinct immune microenvironments.Relative to P2-orange cells, P2-blue cells resided in neighbourhoods enriched for T cells and B cells.P2-blue cells frequently formed clusters around B cell-rich germinal-like centres.P2-orange regions frequently resided inside the lymph node sinuses that were lined by endothelial cells expressing CD34 and PDGFRB.	[41]

Table 5.1: Text for Diffusion-Based Generative Imaging

The text captions were used as the sole input for each pre-trained diffusion model. The training datasets for the models are proprietary. It would be expected that the training images are derived from non-biomedical domains. Outside of the text description, there were no additional cues to guide the image generative process. The results are shown in Figures 5.10 to 5.14.

This adenocarcinoma is arising in a villous adenoma. The surface of the neoplasm is polypoid and reddish pink. Hemorrhage from the surface of the tumor provides for detection with a stool test for occult blood. This neoplasm was located in the sigmoid colon, just out of reach of digital examination, but easily visualized with sigmoidoscopy.

(a)

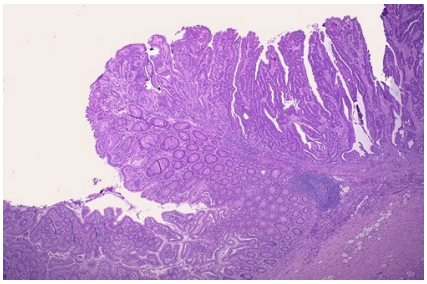


(b) **(c)** **(d)** **(e)**

Figure 5.10: Diffusion generated image of the gross pathology of colorectal cancer. (a) Text embedding for the image (b) Original image (c) DALL.E 2 image (d) Midjourney image (e) Stable Diffusion image

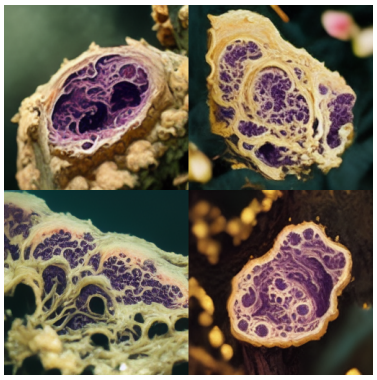
The edge of the carcinoma arising in the villous adenoma is seen here. The neoplastic glands are long and frond-like, similar to those seen in a villous adenoma. The growth is primarily exophytic (outward into the lumen) and invasion is not seen at this point.

(a)

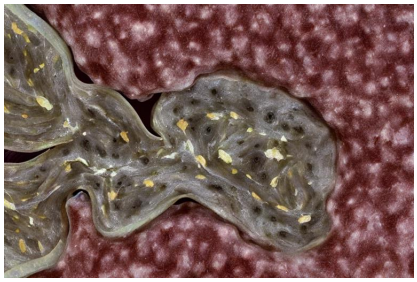


(b)

(c)



(d)

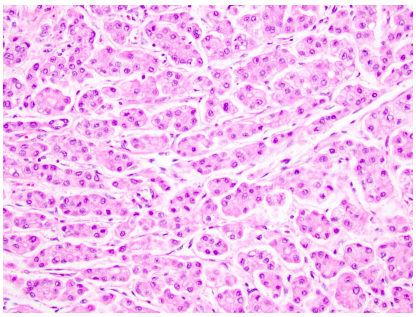


(e)

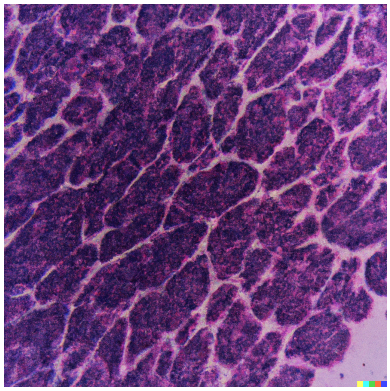
Figure 5.11: Diffusion generated image of the microscopic pathology of colorectal cancer. (a) Text embedding for the image (b) Original image (c) DALL.E 2 image (d) Midjourney image (e) Stable Diffusion image

The tumour cells are polygonal displaying abundant cytoplasm, hyperchromatic enlarged nuclei with high N/C ratio, prominent nucleoli and nuclear irregularity/marked nuclear atypia. The tumour cells are forming nests, pseudo gland formation,trabecular and macro-trabecular patterns.

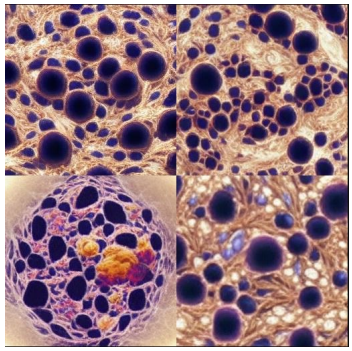
(a)



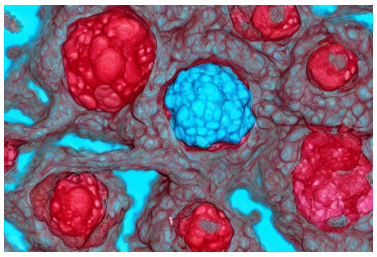
(b)



(c)



(d)

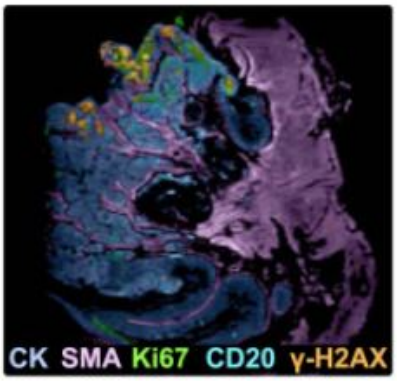


(e)

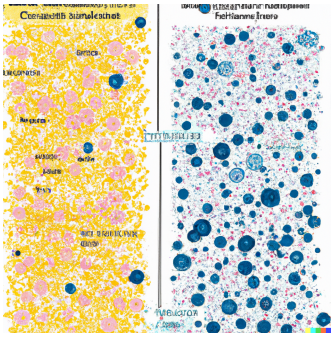
Figure 5.12: Diffusion generated image of the microscopic pathology of hepatocellular cancer. (a) Text embedding for the image (b) Original image (c) DALL.E 2 image (d) Midjourney image (e) Stable Diffusion image

Tumor cells in nascent carcinomas (C-CIA) appear to be highly proliferative, clustering with Ki67 (cluster 1, Fig. 7e), while having weak spatial associations with immune cells (clusters 2), a trend that continues in CRC. Compared to benign adenomas, tumor cells in CRC have significantly weaker spatial associations with cytotoxic T cells, and significantly lower abundances of cytotoxic T cells and TNF-α, suggesting that the anti-tumor response has been suppressed in CRC, which frees tumor cells to divide at higher rates (increased CK/Ki67 spatial association).

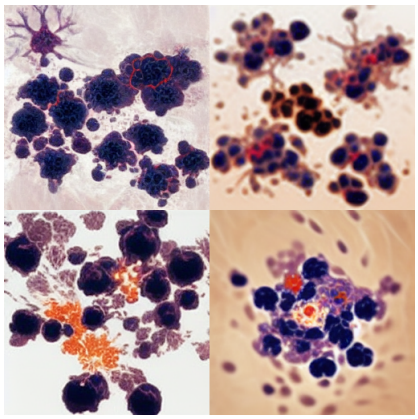
(a)



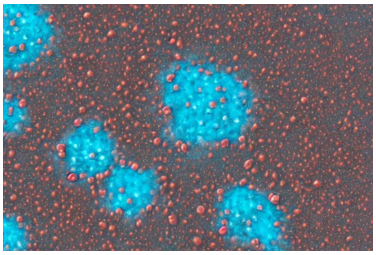
(b)



(c)



(d)

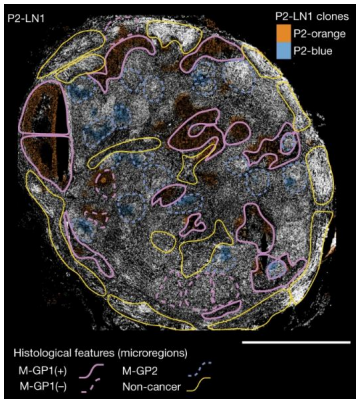


(e)

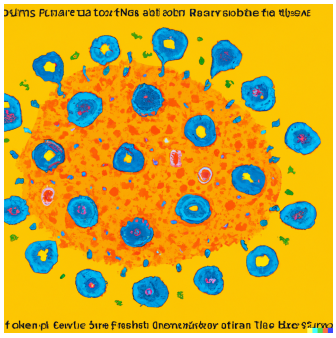
Figure 5.13: Diffusion generated image of tumour-immune dynamics in colorectal cancer. (a) Text embedding for the image (b) Original image (c) DALL.E 2 image (d) Midjourney image (e) Stable Diffusion image

Metastatic subclones occupied distinct immune microenvironments.Relative to P2-orange cells, P2-blue cells resided in neighbourhoods enriched for T cells and B cells.P2-blue cells frequently formed clusters around B cell-rich germinal-like centres.P2-orange regions frequently resided inside the lymph node sinuses that were lined by endothelial cells expressing CD34 and PDGFRB.

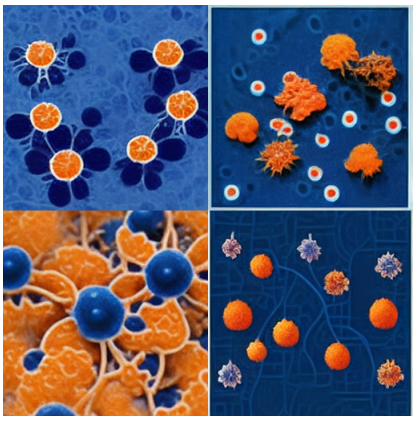
(a)



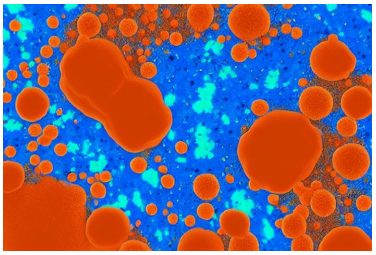
(b)



(c)



(d)



(e)

Figure 5.14: Diffusion generated image of the spatial transcriptomic subclonal architecture in breast cancer. (a) Text embedding for the image (b) Original image (c) DALL.E 2 image (d) Midjourney image (e) Stable Diffusion image

The generated images are surprising. For each image, some features consistent and congruous with the original text are evident. Performance is best with histological descriptions. In particular, image 5.12c, a generated image of the microscopic appearance of hepatocellular carcinoma, shows unexpected similarity with the original image.

The ability to reconstruct even a small set of features in such a specific biomedical context for images which are distinct from standard computer vision training datasets suggests these models have the capacity to learn fundamental principles of nature. This is significant. I believe we are witnessing the early beginnings of a new approach to scientific discovery.

5.3.2.2 Limitations of Diffusion Models

Certainly, the randomly generated images are not faithful recapitulations of the original image. As expected, performance was noticeably sub-optimal when using texts describing detailed tumour-immune spatial relationships from spatial-omics technologies. In their current state, the generated images may be considered as cartoon representations of the real-life image. We are still far away from diffusion models being ready and available for routine use in biology or medicine.

However, I do not believe it is an insurmountable challenge. There is the potential to make large strides with simple measures: adding specific biomedical datasets, eg images from large-scale histology collections, to the training datasets and identifying and developing the optimal text descriptions tailored for use in generative models. In the first instance, I would focus solely on generating H&E images from standard, routinely used pathology descriptions.

Diffusion models do present unique practical challenges. They are formidable models, costing a minimum of £100,000s to train, requiring the latest hardware

and technical support plus leading deep learning expertise. There are a handful of places in the world with the resources to deliver on these requirements. It is an endeavour which would cross traditional academic-commercial and interdisciplinary boundaries.

5.4 Conclusion

In summary, a critical appraisal of currently available multimodal approaches has been presented. The benefits and limitations of WNN, MultiMap and MOFA+ has been discussed in detail. Future perspectives on the use of generative tools, MMVAE and avant-garde diffusion models, for multimodal integration have been considered.

It is a dynamic space. I expect to witness significant advances within the short-term future which may have downstream translational tractability. I am most excited by the insights deep generative modelling tools may be able to offer.

6

Conclusion

The presented body of work describes the application of dissociated single cell sequencing and spatial transcriptomic technologies to clinically-derived breast cancer samples. It is accompanied by a review of the core supporting literature, focusing on the application of machine learning tools in the analysis of high-dimensional datasets and the insights gained from the most contemporary and relevant clinical breast cancer studies.

Droplet-based scRNA-seq, Slide-seq and CARTANA in-situ sequencing were applied to an aggregate collection of five breast cancer samples. Analysis of the scRNA-seq data reveals tumour heterogeneity which is congruous with the published literature.

Reflections on potential implications of the key reviewed literature may offer guidance on avenues for future investigation. The high-dimensional spatial profiling of 700 breast cancer samples from the METABRIC study conducted by Danenberg et al is an important contribution to the field of spatial-omics. It provides a template of multicellular TME structures. The presence of these same TME structures can be investigated across a broader range of tumour types and perturbational clinical settings (e.g. before, on and after treatment). The study also provides guidance on the practical framework and infrastructure required for a large-scale spatial-omics study.

The study by Risom et al applied MIBI and a multiplex antibody panel in a cohort of patients with breast DCIS and stratified samples according to the subsequent development of invasive breast cancer. The findings of the study suggest that a thin myoepithelium in DCIS is less frequently associated with the future development of invasive breast cancer. The first step would be to reproduce this observation across a wider cohort of clinical DCIS samples from a range of clinical sources across a panel of complementary spatial-omics technologies. Furthermore, in such a context, a comparison could be conducted between the use of Ki-67 and the multivariate score of global proliferation (MPI score)

as proposed by Gaglia et al. Previous work has shown Ki-67 to be a predictive marker for recurrence in breast DCIS. Therefore, it would be constructive to explore for improved predictive performance with the MPI score.

If the findings remain reproducible, it would then be constructive to (1) explore the mechanistic factors underpinning this observation and (2) investigate the relationship between myoepithelial thickness and detailed clinical relapse/survival outcomes. Conditioned on reproducibility and detailed mechanistic insight, the longer-term aim may be to develop myoepithelial depth as a biomarker for stratification of adjuvant therapy.

Detailed evaluation of commonly used algorithms used in the analysis of scRNA-seq data suggests caution in utilising methods which are based on precisely-defined high-dimensional vector spaces, spaces which have characteristic a priori types of 'behaviour'. Different types of data 'live' in different types of vector spaces which results in different data properties. Dataset-algorithmic pairing should be based on the properties of data under exploration. Shotgun algorithmic pairing is best avoided.

A novel component of the thesis is the application of diffusion models on histopathology text descriptions for text-image translation. The experimental results presented are unidirectional but there is certainly scope for bidirectional translation (i.e. text-image translation and image-text translation). Diffusion models are a rapidly developing tool in the field of generative deep learning tools. They have shown potential for more traditional applications of image generation. Their utility in a biomedical context remains as yet to be determined.

Questions remain surrounding the challenges in developing the infrastructure to develop scalable machine learning tools for clinical use which is accompanied by the requirement for data collection, curation and harmonisation within and across different IT systems. Furthermore, all such progress is to be conducted whilst maintaining ethical practices which respect the diversity

inherent to clinical datasets and safeguard data confidentiality.

The marriage between medicine and machine learning has just begun. It is a pivotal moment. As a field, we have the opportunity to establish a strong framework which is guided by robust practices and shaped by factors specific to the nuances and types of bias common to biomedical datasets. Decisions made now on how the frameworks will be designed and operate have the potential to contribute to constructive advance in the future. It is an exciting opportunity to be part of and witness the reshaping of our field. The call for action has been made by the field.

Bibliography

[1] Indu Agarwal and Luis Blanco. Pathology of the breast, 2022. https://www.pathologyoutlines.com/topic/breastnormal.html, Last accessed on 2022-12-17.

[2] Mona Shehata, Andrew Teschendorff, Gemma Sharp, Nikola Novcic, I Alasdair Russell, Stefanie Avril, Michael Prater, Peter Eirew, Carlos Caldas, Christine J Watson, and John Stingl. Phenotypic and functional characterisation of the luminal cell hierarchy of the mammary gland. *Breast Cancer Res.*, 14(5):R134, October 2012.

[3] John Stingl, Peter Eirew, Ian Ricketson, Mark Shackleton, François Vaillant, David Choi, Haiyan I Li, and Connie J Eaves. Purification and unique properties of mammary epithelial stem cells. *Nature*, 439(7079):993–997, February 2006.

[4] Quy H Nguyen, Nicholas Pervolarakis, Kerrigan Blake, Dennis Ma, Ryan Tevia Davis, Nathan James, Anh T Phung, Elizabeth Willey, Raj Kumar, Eric Jabart, Ian Driver, Jason Rock, Andrei Goga, Seema A Khan, Devon A Lawson, Zena Werb, and Kai Kessenbrock. Profiling human breast epithelial cells using single cell RNA sequencing identifies cell diversity. *Nat. Commun.*, 9(1):2028, May 2018.

[5] Monica Brown, Alex Tsodikov, Katrina R Bauer, Carol A Parise, and Vincent Caggiano. The role of human epidermal growth factor receptor 2 in the

survival of women with estrogen and progesterone receptor-negative, invasive breast cancer: the california cancer registry, 1999-2004. *Cancer*, 112(4):737–747, February 2008.

[6] Bruce G Haffty, Qifeng Yang, Michael Reiss, Thomas Kearney, Susan A Higgins, Joanne Weidhaas, Lyndsay Harris, Willam Hait, and Deborah Toppmeyer. Locoregional relapse and distant metastasis in conservatively managed triple negative early-stage breast cancer. *J. Clin. Oncol.*, 24(36):5652–5657, December 2006.

[7] Ju-Yi Hsu, Chee-Jen Chang, and Jur-Shan Cheng. Survival, treatment regimens and medical costs of women newly diagnosed with metastatic triple-negative breast cancer. *Sci. Rep.*, 12(1):729, January 2022.

[8] Felipe C Geyer, Fresia Pareja, Britta Weigelt, Emad Rakha, Ian O Ellis, Stuart J Schnitt, and Jorge S Reis-Filho. The spectrum of triple-negative breast disease: High- and low-grade lesions. *Am. J. Pathol.*, 187(10):2139–2151, October 2017.

[9] C M Perou, T Sørlie, M B Eisen, M van de Rijn, S S Jeffrey, C A Rees, J R Pollack, D T Ross, H Johnsen, L A Akslen, O Fluge, A Pergamenschikov, C Williams, S X Zhu, P E Lønning, A L Børresen-Dale, P O Brown, and D Botstein. Molecular portraits of human breast tumours. *Nature*, 406(6797):747–752, August 2000.

[10] H Raza Ali, Oscar M Rueda, Suet-Feung Chin, Christina Curtis, Mark J Dunning, Samuel Ajr Aparicio, and Carlos Caldas. Genome-driven integrated classification of breast cancer validated in over 7,500 samples. *Genome Biol.*, 15(8):431, August 2014.

[11] Ana C Garrido-Castro, Nancy U Lin, and Kornelia Polyak. Insights into molecular classifications of triple-negative breast cancer: Improving

patient selection for treatment. *Cancer Discov.*, 9(2):176–198, February 2019.

[12] A Prat, A Lluch, J Albanell, W T Barry, C Fan, J I Chacón, J S Parker, L Calvo, A Plazaola, A Arcusa, M A Seguí-Palmer, O Burgues, N Ribelles, A Rodriguez-Lescure, A Guerrero, M Ruiz-Borrego, B Munarriz, J A López, B Adamo, M C U Cheang, Y Li, Z Hu, M L Gulley, M J Vidal, B N Pitcher, M C Liu, M L Citron, M J Ellis, E Mardis, T Vickery, C A Hudis, E P Winer, L A Carey, R Caballero, E Carrasco, M Martín, C M Perou, and E Alba. Predicting response and survival in chemotherapy-treated triple-negative breast cancer. *Br. J. Cancer*, 111(8):1532–1541, October 2014.

[13] Brian D Lehmann, Joshua A Bauer, Xi Chen, Melinda E Sanders, A Bapsi Chakravarthy, Yu Shyr, and Jennifer A Pietenpol. Identification of human triple-negative breast cancer subtypes and preclinical models for selection of targeted therapies. *J. Clin. Invest.*, 121(7):2750–2767, July 2011.

[14] Andrew Tutt, Holly Tovey, Maggie Chon U Cheang, Sarah Kernaghan, Lucy Kilburn, Patrycja Gazinska, Julie Owen, Jacinta Abraham, Sophie Barrett, Peter Barrett-Lee, Robert Brown, Stephen Chan, Mitchell Dowsett, James M Flanagan, Lisa Fox, Anita Grigoriadis, Alexander Gutin, Catherine Harper-Wynne, Matthew Q Hatton, Katherine A Hoadley, Jyoti Parikh, Peter Parker, Charles M Perou, Rebecca Roylance, Vandna Shah, Adam Shaw, Ian E Smith, Kirsten M Timms, Andrew M Wardley, Gregory Wilson, Cheryl Gillett, Jerry S Lanchbury, Alan Ashworth, Nazneen Rahman, Mark Harries, Paul Ellis, Sarah E Pinder, and Judith M Bliss. Carboplatin in BRCA1/2-mutated and triple-negative breast cancer BRCAness subgroups: the TNT trial. *Nat. Med.*, 24(5):628–637, May 2018.

[15] Hiroko Masuda, Keith A Baggerly, Ying Wang, Ya Zhang, Ana Maria Gonzalez-Angulo, Funda Meric-Bernstam, Vicente Valero, Brian D Lehmann, Jennifer A Pietenpol, Gabriel N Hortobagyi, W Fraser Symmans, and Naoto T Ueno. Differential response to neoadjuvant chemotherapy among 7 triple-negative breast cancer molecular subtypes. *Clin. Cancer Res.*, 19(19):5533–5540, October 2013.

[16] Mihriban Karaayvaz, Simona Cristea, Shawn M Gillespie, Anoop P Patel, Ravindra Mylvaganam, Christina C Luo, Michelle C Specht, Bradley E Bernstein, Franziska Michor, and Leif W Ellisen. Unravelling subclonal heterogeneity and aggressive disease states in TNBC through single-cell RNA-seq. *Nat. Commun.*, 9(1):3588, September 2018.

[17] Laura J van 't Veer, Hongyue Dai, Marc J van de Vijver, Yudong D He, Augustinus A M Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, George J Schreiber, Ron M Kerkhoven, Chris Roberts, Peter S Linsley, René Bernards, and Stephen H Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, January 2002.

[18] Fatima Cardoso, Laura J van't Veer, Jan Bogaerts, Leen Slaets, Giuseppe Viale, Suzette Delaloge, Jean-Yves Pierga, Etienne Brain, Sylvain Causeret, Mauro DeLorenzi, Annuska M Glas, Vassilis Golfinopoulos, Theodora Goulioti, Susan Knox, Erika Matos, Bart Meulemans, Peter A Neijenhuis, Ulrike Nitz, Rodolfo Passalacqua, Peter Ravdin, Isabel T Rubio, Mahasti Saghatchian, Tineke J Smilde, Christos Sotiriou, Lisette Stork, Carolyn Straehle, Geraldine Thomas, Alastair M Thompson, Jacobus M van der Hoeven, Peter Vuylsteke, René Bernards, Konstantinos Tryfonidis, Emiel Rutgers, Martine Piccart, and MINDACT Investigators. 70-gene signature

as an aid to treatment decisions in early-stage breast cancer. *N. Engl. J. Med.*, 375(8):717–729, August 2016.

[19] Devon A Lawson, Nirav R Bhakta, Kai Kessenbrock, Karin D Prummel, Ying Yu, Ken Takai, Alicia Zhou, Henok Eyob, Sanjeev Balakrishnan, Chih-Yang Wang, Paul Yaswen, Andrei Goga, and Zena Werb. Single-cell analysis reveals a stem-cell program in human metastatic breast cancer cells. *Nature*, 526(7571):131–135, October 2015.

[20] Giovanni Ciriello, Martin L Miller, Bülent Arman Aksoy, Yasin Senbabaoglu, Nikolaus Schultz, and Chris Sander. Emerging landscape of oncogenic signatures across human cancers. *Nat. Genet.*, 45(10):1127–1133, October 2013.

[21] Hartland W Jackson, Jana R Fischer, Vito R T Zanotelli, H Raza Ali, Robert Mechera, Savas D Soysal, Holger Moch, Simone Muenst, Zsuzsanna Varga, Walter P Weber, and Bernd Bodenmiller. The single-cell pathology landscape of breast cancer. *Nature*, 578(7796):615–620, February 2020.

[22] H Raza Ali, Hartland W Jackson, Vito R T Zanotelli, Esther Danenberg, Jana R Fischer, Helen Bardwell, Elena Provenzano, CRUK IMAXT Grand Challenge Team, Oscar M Rueda, Suet-Feung Chin, Samuel Aparicio, Carlos Caldas, and Bernd Bodenmiller. Imaging mass cytometry and multiplatform genomics define the phenogenomic landscape of breast cancer. *Nat Cancer*, 1(2):163–175, February 2020.

[23] Esther Danenberg, Helen Bardwell, Vito R T Zanotelli, Elena Provenzano, Suet-Feung Chin, Oscar M Rueda, Andrew Green, Emad Rakha, Samuel Aparicio, Ian O Ellis, Bernd Bodenmiller, Carlos Caldas, and H Raza Ali. Breast tumor microenvironment structures are associated with genomic features and clinical outcome. *Nat. Genet.*, 54(5):660–669, May 2022.

[24] Jan Kåhre. *General Properties of Information*, pages 41–79. Springer US, Boston, MA, 2002.

[25] Marcel Smid, F Germán Rodríguez-González, Anieta M Sieuwerts, Roberto Salgado, Wendy J C Prager-Van der Smissen, Michelle van der Vlugt-Daane, Anne van Galen, Serena Nik-Zainal, Johan Staaf, Arie B Brinkman, Marc J van de Vijver, Andrea L Richardson, Aquila Fatima, Kim Berentsen, Adam Butler, Sancha Martin, Helen R Davies, Reno Debets, Marion E Meijer-Van Gelder, Carolien H M van Deurzen, Gaëtan MacGrogan, Gert G G M Van den Eynden, Colin Purdie, Alastair M Thompson, Carlos Caldas, Paul N Span, Peter T Simpson, Sunil R Lakhani, Steven Van Laere, Christine Desmedt, Markus Ringnér, Stefania Tommasi, Jorunn Eyford, Annegien Broeks, Anne Vincent-Salomon, P Andrew Futreal, Stian Knappskog, Tari King, Gilles Thomas, Alain Viari, Anita Langerød, Anne-Lise Børresen-Dale, Ewan Birney, Hendrik G Stunnenberg, Mike Stratton, John A Foekens, and John W M Martens. Breast cancer genome and transcriptome integration implicates specific mutational signatures with immune cell infiltration. *Nat. Commun.*, 7(1):12910, September 2016.

[26] Michael S Rooney, Sachet A Shukla, Catherine J Wu, Gad Getz, and Nir Hacohen. Molecular and genetic properties of tumors associated with local immune cytolytic activity. *Cell*, 160(1-2):48–61, January 2015.

[27] H Raza Ali, Leon Chlon, Paul D P Pharoah, Florian Markowetz, and Carlos Caldas. Patterns of immune infiltration in breast cancer and their clinical implications: A gene-expression-based retrospective study. *PLoS Med.*, 13(12):e1002194, December 2016.

[28] Stephen-John Sammut, Mireia Crispin-Ortuzar, Suet-Feung Chin, Elena Provenzano, Helen A Bardwell, Wenxin Ma, Wei Cope, Ali Dariush, Sarah-

Jane Dawson, Jean E Abraham, Janet Dunn, Louise Hiller, Jeremy Thomas, David A Cameron, John M S Bartlett, Larry Hayward, Paul D Pharoah, Florian Markowetz, Oscar M Rueda, Helena M Earl, and Carlos Caldas. Multi-omic machine learning predictor of breast cancer therapy response. *Nature*, 601(7894):623–629, January 2022.

[29] Susanne Nichterwitz, Julio Aguila Benitez, Rein Hoogstraaten, Qiaolin Deng, and Eva Hedlund. Lcm-seq: A method for spatial transcriptomic profiling using laser capture microdissection coupled with polya-based rna sequencing. *Methods in molecular biology (Clifton, N.J.)*, 1649:95–110, 2018.

[30] Karin Schütze and Georgia Lahr. Laser micromanipulation systems as universal tools in cellular and molecular biology and in medicine. *Cellular and Molecular Biology*, 5, 1998.

[31] Je Hyuk Lee, Evan R Daugharthy, Jonathan Scheiman, Reza Kalhor, Thomas C Ferrante, Richard Terry, Brian M Turczyk, Joyce L Yang, Ho Suk Lee, John Aach, Kun Zhang, and George M Church. Fluorescent in situ sequencing (fisseq) of rna for gene expression profiling in intact cells and tissues. *Nature Protocols 2015 10:3*, 10:442–458, 2 2015.

[32] Jessica Svedlund, Carina Strell, Xiaoyan Qian, Kilian J.C. Zilkens, Nicholas P. Tobin, Jonas Bergh, Anieta M. Sieuwerts, and Mats Nilsson. Generation of in situ sequencing based oncomaps to spatially resolve gene expression profiles of diagnostic and prognostic markers in breast cancer. *EBioMedicine*, 48:212–223, 10 2019.

[33] Patrik L. Ståhl, Fredrik Salmén, Sanja Vickovic, Anna Lundmark, José Fernández Navarro, Jens Magnusson, Stefania Giacomello, Michaela Asp, Jakub O. Westholm, Mikael Huss, Annelie Mollbrink, Sten Linnarsson,

Simone Codeluppi, Åke Borg, Fredrik Pontén, Paul Igor Costea, Pelin Sahlén, Jan Mulder, Olaf Bergmann, Joakim Lundeberg, and Jonas Frisén. Visualization and analysis of gene expression in tissue sections by spatial transcriptomics. *Science*, 353:78–82, 7 2016.

[34] Samuel G. Rodriques, Robert R. Stickels, Aleksandrina Goeva, Carly A. Martin, Evan Murray, Charles R. Vanderburg, Joshua Welch, Linlin M. Chen, Fei Chen, and Evan Z. Macosko. Slide-seq: A scalable technology for measuring genome-wide expression at high spatial resolution. *Science*, 363(6434):1463–1467, mar 2019.

[35] Sanja Vickovic, Gökcen Eraslan, Fredrik Salmén, Johanna Klughammer, Linnea Stenbeck, Denis Schapiro, Tarmo Äijö, Richard Bonneau, Ludvig Bergenstråhle, José Fernandéz Navarro, Joshua Gould, Gabriel K. Griffin, Åke Borg, Mostafa Ronaghi, Jonas Frisén, Joakim Lundeberg, Aviv Regev, and Patrik L. Ståhl. High-definition spatial transcriptomics for in situ tissue profiling. *Nature Methods 2019 16:10*, 16:987–990, 9 2019.

[36] Ilana Schlam, Sarah E. Church, Tyler D. Hether, Krysta Chaldekas, Briana M. Hudson, Andrew M. White, Emily Maisonet, Brent T. Harris, and Sandra M. Swain. The tumor immune microenvironment of primary and metastatic her2- positive breast cancers utilizing gene expression and spatial proteomic profiling. *Journal of Translational Medicine*, 19:480, 12 2021.

[37] Ilya Korsunsky, Nghia Millard, Jean Fan, Kamil Slowikowski, Fan Zhang, Kevin Wei, Yuriy Baglaenko, Michael Brenner, Po ru Loh, and Soumya Raychaudhuri. Fast, sensitive and accurate integration of single-cell data with harmony. *Nature Methods 2019 16:12*, 16:1289–1296, 11 2019.

[38] Vivien Marx. Method of the year: spatially resolved transcriptomics. *Nature Methods 2021 18:1*, 18:9–14, 1 2021.

[39] Andrew Erickson, Mengxiao He, Emelie Berglund, Maja Marklund, Reza Mirzazadeh, Niklas Schultz, Linda Kvastad, Alma Andersson, Ludvig Bergenstråhle, Joseph Bergenstråhle, Ludvig Larsson, Leire Alonso Galicia, Alia Shamikh, Elisa Basmaci, Teresita Díaz De Ståhl, Timothy Rajakumar, Dimitrios Doultsinos, Kim Thrane, Andrew L Ji, Paul A Khavari, Firaz Tarish, Anna Tanoglidi, Jonas Maaskola, Richard Colling, Tuomas Mirtti, Freddie C Hamdy, Dan J Woodcock, Thomas Helleday, Ian G Mills, Alastair D Lamb, and Joakim Lundeberg. Spatially resolved clonal copy number alterations in benign and malignant tissue. *Nature*, 608(7922):360–367, August 2022.

[40] William L Hwang, Karthik A Jagadeesh, Jimmy A Guo, Hannah I Hoffman, Payman Yadollahpour, Jason W Reeves, Rahul Mohan, Eugene Drokhlyansky, Nicholas Van Wittenberghe, Orr Ashenberg, Samouil L Farhi, Denis Schapiro, Prajan Divakar, Eric Miller, Daniel R Zollinger, George Eng, Jason M Schenkel, Jennifer Su, Carina Shiau, Patrick Yu, William A Freed-Pastor, Domenic Abbondanza, Arnav Mehta, Joshua Gould, Conner Lambden, Caroline B M Porter, Alexander Tsankov, Danielle Dionne, Julia Waldman, Michael S Cuoco, Lan Nguyen, Toni Delorey, Devan Phillips, Jaimie L Barth, Marina Kem, Clifton Rodrigues, Debora Ciprani, Jorge Roldan, Piotr Zelga, Vjola Jorgji, Jonathan H Chen, Zackery Ely, Daniel Zhao, Kit Fuhrman, Robin Fropf, Joseph M Beechem, Jay S Loeffler, David P Ryan, Colin D Weekes, Cristina R Ferrone, Motaz Qadan, Martin J Aryee, Rakesh K Jain, Donna S Neuberg, Jennifer Y Wo, Theodore S Hong, Ramnik Xavier, Andrew J Aguirre, Orit Rozenblatt-Rosen, Mari Mino-Kenudson, Carlos Fernandez-Del Castillo, Andrew S Liss, David T Ting, Tyler Jacks, and Aviv

Regev. Single-nucleus and spatial transcriptome profiling of pancreatic cancer identifies multicellular dynamics associated with neoadjuvant treatment. *Nat. Genet.*, 54(8):1178–1191, August 2022.

[41] Artem Lomakin, Jessica Svedlund, Carina Strell, Milana Gataric, Artem Shmatko, Gleb Rukhovich, Jun Sung Park, Young Seok Ju, Stefan Dentro, Vitalii Kleshchevnikov, Vasyl Vaskivskyi, Tong Li, Omer Ali Bayraktar, Sarah Pinder, Andrea L Richardson, Sandro Santagata, Peter J Campbell, Hege Russnes, Moritz Gerstung, Mats Nilsson, and Lucy R Yates. Spatial genomics maps the structure, nature and evolution of cancer clones. *Nature*, 611(7936):594–602, November 2022.

[42] Luyi Tian, Fei Chen, and Evan Z. Macosko. The expanding vistas of spatial transcriptomics. *Nature Biotechnology*, 2022.

[43] Giorgio Gaglia, Sheheryar Kabraji, Danae Rammos, Yang Dai, Ana Verma, Shu Wang, Caitlin E Mills, Mirra Chung, Johann S Bergholz, Shannon Coy, Jia-Ren Lin, Rinath Jeselsohn, Otto Metzger, Eric P Winer, Deborah A Dillon, Jean J Zhao, Peter K Sorger, and Sandro Santagata. Temporal and spatial topography of cell proliferation in cancer. *Nat. Cell Biol.*, 24(3):316–326, March 2022.

[44] Torsten O Nielsen, Samuel C Y Leung, David L Rimm, Andrew Dodson, Balazs Acs, Sunil Badve, Carsten Denkert, Matthew J Ellis, Susan Fineberg, Margaret Flowers, Hans H Kreipe, Anne-Vibeke Laenkholm, Hongchao Pan, Frédérique M Penault-Llorca, Mei-Yin Polley, Roberto Salgado, Ian E Smith, Tomoharu Sugie, John M S Bartlett, Lisa M McShane, Mitch Dowsett, and Daniel F Hayes. Assessment of ki67 in breast cancer: Updated recommendations from the international ki67 in breast cancer working group. *J. Natl. Cancer Inst.*, 113(7):808–819, July 2021.

[45] Tyler Risom, David R Glass, Inna Averbukh, Candace C Liu, Alex Baranski, Adam Kagel, Erin F McCaffrey, Noah F Greenwald, Belén Rivero-Gutiérrez, Siri H Strand, Sushama Varma, Alex Kong, Leeat Keren, Sucheta Srivastava, Chunfang Zhu, Zumana Khair, Deborah J Veis, Katherine Deschryver, Sujay Vennam, Carlo Maley, E Shelley Hwang, Jeffrey R Marks, Sean C Bendall, Graham A Colditz, Robert B West, and Michael Angelo. Transition to invasive breast cancer is associated with progressive changes in the structure and composition of tumor stroma. *Cell*, 185(2):299–310.e18, January 2022.

[46] Dylan M Cable, Evan Murray, Luli S Zou, Aleksandrina Goeva, Evan Z Macosko, Fei Chen, and Rafael A Irizarry. Robust decomposition of cell type mixtures in spatial transcriptomics. *Nat. Biotechnol.*, 40(4):517–526, April 2022.

[47] Romain Lopez, Achille Nazaret, Maxime Langevin, Jules Samaran, Jeffrey Regier, Michael I. Jordan, and Nir Yosef. A joint model of unpaired data from scrna-seq and spatial transcriptomics for imputing missing gene expression measurements. *pre-print*, 2019.

[48] Jian Hu, Xiangjie Li, Kyle Coleman, Amelia Schroeder, Nan Ma, David J Irwin, Edward B Lee, Russell T Shinohara, and Mingyao Li. SpaGCN: Integrating gene expression, spatial location and histology to identify spatial domains and spatially variable genes by graph convolutional network. *Nat. Methods*, 18(11):1342–1351, November 2021.

[49] Edward Zhao, Matthew R Stone, Xing Ren, Jamie Guenthoer, Kimberly S Smythe, Thomas Pulliam, Stephen R Williams, Cedric R Uytingco, Sarah E B Taylor, Paul Nghiem, Jason H Bielas, and Raphael Gottardo. Spatial

transcriptomics at subspot resolution with BayesSpace. *Nat. Biotechnol.*, 39(11):1375–1384, November 2021.

[50] Zixuan Cang and Qing Nie. Inferring spatial and signaling relationships between cells from single cell transcriptomic data. *Nat. Commun.*, 11(1):2084, April 2020.

[51] Simon "Graham, Quoc Dang Vu, Shan E Ahmed Raza, Ayesha Azam, Yee Wah Tsang, Jin Tae Kwak, and Nasir" Rajpoot. "hover-net: Simultaneous segmentation and classification of nuclei in multi-tissue histology images", 2018.

[52] Alexey "Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil" Houlsby. "an image is worth 16x16 words: Transformers for image recognition at scale", 2020.

[53] NYU Langone Medical Center Experimental Pathology Research Laboratory. Freezing tissues for histology: Embedding protocol for Frozen Samples.

[54] Taghreed Hirz, Shenglin Mei, Hirak Sarkar, Youmna Kfoury, Shulin Wu, Bronte M. Verhoeven, Alexander O. Subtelny, Dimitar V. Zlatev, Matthew W. Wszolek, Keyan Salari, Evan Murray, Fei Chen, Evan Z. Macosko, Chin-Lee Wu, David T. Scadden, Douglas M. Dahl, Ninib Baryawno, Philip J. Saylor, Peter V. Kharchenko, and David B. Sykes. Integrated single-cell and spatial transcriptomic analyses unravel the heterogeneity of the prostate tumor microenvironment. *bioRxiv*, 2022.

[55] EF Gaffney, PH Riegman, WE Grizzle, and PH Watson. Factors that drive the increasing use of ffpe tissue in basic and translational cancer research. *https://doi.org/10.1080/10520295.2018.1446101*, 93:373–386, 7 2018.

[56] Elena Favaro, Simon Lord, Adrian L Harris, and Francesca M Buffa. Gene expression and hypoxia in breast cancer. *Genome Medicine 2011 3:8*, 3:1–12, 8 2011.

[57] JEUNG IL KIM, KYUNG UN CHOI, IN SOOK LEE, YOUNG JIN CHOI, WON TACK KIM, DONG HOON SHIN, KYUNGBIN KIM, JEONG HEE LEE, JEE YEON KIM, and MEE YOUNG SOL. Expression of hypoxic markers and their prognostic significance in soft tissue sarcoma. *Oncology Letters*, 9:1699, 2015.

[58] Brita S. Sørensen, Jing Hao, Jens Overgaard, Jan Alsner, and Michael R. Horsman. Validating the use of caix and glut1 as endogenous markers for hypoxia in solid tumors. *Cancer Research*, 65, 2005.

[59] Woosung Chung, Hye Hyeon Eum, Hae Ock Lee, Kyung Min Lee, Han Byoel Lee, Kyu Tae Kim, Han Suk Ryu, Sangmin Kim, Jeong Eon Lee, Yeon Hee Park, Zhengyan Kan, Wonshik Han, and Woong Yang Park. Single-cell RNA-seq enables comprehensive tumour and immune cell profiling in primary breast cancer. *Nature Communications*, 8(1):1–12, may 2017.

[60] F. M. Buffa, A. L. Harris, C. M. West, and C. J. Miller. Large meta-analysis of multiple cancers reveals a common, compact and highly prognostic hypoxia metagene. *British Journal of Cancer 2010 102:2*, 102:428–435, 1 2010.

[61] Evan Z. Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R. Bialas, Nolan Kamitaki,

Emily M. Martersteck, John J. Trombetta, David A. Weitz, Joshua R. Sanes, Alex K. Shalek, Aviv Regev, and Steven A. McCarroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161:1202, 5 2015.

[62] Shannon Axelrod, Matthew Cai, Ambrose Carr, Jeremy Freeman, Deep Ganguli, Justin Kiggins, Brian Long, Tony Tung, and Kevin Yamauchi. starfish: scalable pipelines for image-based transcriptomics. *Journal of Open Source Software*, 6:2440, 05 2021.

[63] Tommaso Biancalani, Gabriele Scalia, Lorenzo Buffoni, Raghav Avasthi, Ziqing Lu, Aman Sanger, Neriman Tokcan, Charles R Vanderburg, Åsa Segerstolpe, Meng Zhang, Inbal Avraham-Davidi, Sanja Vickovic, Mor Nitzan, Sai Ma, Ayshwarya Subramanian, Michal Lipinski, Jason Buenrostro, Nik Bear Brown, Duccio Fanelli, Xiaowei Zhuang, Evan Z Macosko, and Aviv Regev. Deep learning and alignment of spatially resolved single-cell transcriptomes with tangram. *Nat. Methods*, 18(11):1352–1362, November 2021.

[64] Sunny Z. Wu, Ghamdan Al-Eryani, Daniel Lee Roden, Simon Junankar, Kate Harvey, Alma Andersson, Aatish Thennavan, Chenfei Wang, James R. Torpy, Nenad Bartonicek, Taopeng Wang, Ludvig Larsson, Dominik Kaczorowski, Neil I. Weisenfeld, Cedric R. Uytingco, Jennifer G. Chew, Zachary W. Bent, Chia Ling Chan, Vikkitharan Gnanasambandapillai, Charles Antoine Dutertre, Laurence Gluch, Mun N. Hui, Jane Beith, Andrew Parker, Elizabeth Robbins, Davendra Segara, Caroline Cooper, Cindy Mak, Belinda Chan, Sanjay Warrier, Florent Ginhoux, Ewan Millar, Joseph E. Powell, Stephen R. Williams, X. Shirley Liu, Sandra O'Toole, Elgene Lim, Joakim Lundeberg, Charles M. Perou, and Alexander Swarbrick.

A single-cell and spatially resolved atlas of human breast cancers. *Nature Genetics 2021 53:9*, 53:1334–1347, 9 2021.

[65] Malte D Luecken and Fabian J Theis. Current best practices in single-cell rna-seq analysis: a tutorial. *Molecular Systems Biology*, 15:e8746, 6 2019.

[66] Koki Tsuyuzaki, Hiroyuki Sato, Kenta Sato, and Itoshi Nikaido. Benchmarking principal component analysis for large-scale single-cell rna-sequencing. *Genome Biology*, 21:1–17, 1 2020.

[67] Jake Lever, Martin Krzywinski, and Naomi Altman. Points of significance: Principal component analysis. *Nature Methods*, 14:641–642, 6 2017.

[68] Mats Björklund and Bj¨ Björklund. Be careful with your principal components. *Evolution*, 73:2151–2158, 10 2019.

[69] Zijing Liu and Mauricio Barahona. Graph-based data clustering via multiscale community detection. *Applied Network Science*, 5:1–20, 12 2020.

[70] Yael Baran, Akhiad Bercovich, Arnau Sebe-Pedros, Yaniv Lubling, Amir Giladi, Elad Chomsky, Zohar Meir, Michael Hoichman, Aviezer Lifshitz, and Amos Tanay. Metacell: Analysis of single-cell rna-seq data using k-nn graph partitions. *Genome Biology*, 20:1–19, 10 2019.

[71] Wan-Lei Zhao, Hui Wang, and Chong-Wah Ngo. Approximate k-nn graph construction: a generic online approach, 2018.

[72] V B Surya Prasath, Haneen Arafat Abu Alfeilat, Ahmad B.A. Hassanat, Omar Lasassmeh, Ahmad S. Tarawneh, Mahmoud Bashir Alhasanat, Hamzeh S. Eyal Salman, and V.B. Surya Prasath. Effects of distance measure choice on k-nearest neighbor classifier performance: A review. *Big Data*, 7(4):221–248, dec 2019.

[73] Marcello D'agostino, Valentino Dardanoni, ' Agostino, and V Dardanoni. What's so special about euclidean distance? *Social Choice and Welfare 2008 33:2*, 33:211–233, 12 2008.

[74] Sayan Ghosh, Mahantesh Halappanavar, Antonino Tumeo, and Ananth Kalyanarainan. Scaling and quality of modularity optimization methods for graph clustering. *2019 IEEE High Performance Extreme Computing Conference, HPEC 2019*, 9 2019.

[75] Vincent D. Blondel, Jean-Loup Guillaume, Renaud Lambiotte, and Etienne Lefebvre. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2008, 3 2008.

[76] Etienne Becht, Leland McInnes, John Healy, Charles Antoine Dutertre, Immanuel W.H. Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology 2018 37:1*, 37:38–44, 12 2018.

[77] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 2 2018.

[78] Allen Hatcher. *Algebraic Topology*. Cambridge University Press, 10 2009.

[79] Open cover – from wolfram mathworld.

[80] Ulrich Bauer, Michael Kerber, Fabian Roll, and Alexander Rolle. A unified view on the functorial nerve theorem and its variations. *arXiv*, 2022.

[81] Christina Curtis, Sohrab P. Shah, Suet Feung Chin, Gulisa Turashvili, Oscar M. Rueda, Mark J. Dunning, Doug Speed, Andy G. Lynch, Shamith Samarajiwa, Yinyin Yuan, Stefan Gräf, Gavin Ha, Gholamreza Haffari, Ali Bashashati, Roslin Russell, Steven McKinney, Samuel Aparicio, James D.

Brenton, Ian Ellis, David Huntsman, Sarah Pinder, Leigh Murphy, Helen Bardwell, Zhihao Ding, Linda Jones, Bin Liu, Irene Papatheodorou, Stephen J. Sammut, Gordon Wishart, Steven Chia, Karen Gelmon, Caroline Speers, Peter Watson, Roger Blamey, Andrew Green, Douglas MacMillan, Emad Rakha, Cheryl Gillett, Anita Grigoriadis, Emanuele De Rinaldis, Andy Tutt, Michelle Parisien, Sandra Troup, Derek Chan, Claire Fielding, Ana Teresa Maia, Sarah McGuire, Michelle Osborne, Sara M. Sayalero, Inmaculada Spiteri, James Hadfield, Lynda Bell, Katie Chow, Nadia Gale, Maria Kovalik, Ying Ng, Leah Prentice, Simon Tavaré, Florian Markowetz, Anita Langerød, Elena Provenzano, Arnie Purushotham, Anne Lise Børresen-Dale, and Carlos Caldas. The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature 2012 486:7403*, 486:346–352, 4 2012.

[82] Sohrab P. Shah, Andrew Roth, Rodrigo Goya, Arusha Oloumi, Gavin Ha, Yongjun Zhao, Gulisa Turashvili, Jiarui Ding, Kane Tse, Gholamreza Haffari, Ali Bashashati, Leah M. Prentice, Jaswinder Khattra, Angela Burleigh, Damian Yap, Virginie Bernard, Andrew McPherson, Karey Shumansky, Anamaria Crisan, Ryan Giuliany, Alireza Heravi-Moussavi, Jamie Rosner, Daniel Lai, Inanc Birol, Richard Varhol, Angela Tam, Noreen Dhalla, Thomas Zeng, Kevin Ma, Simon K. Chan, Malachi Griffith, Annie Moradian, S. W.Grace Cheng, Gregg B. Morin, Peter Watson, Karen Gelmon, Stephen Chia, Suet Feung Chin, Christina Curtis, Oscar M. Rueda, Paul D. Pharoah, Sambasivarao Damaraju, John MacKey, Kelly Hoon, Timothy Harkins, Vasisht Tadigotla, Mahvash Sigaroudinia, Philippe Gascard, Thea Tlsty, Joseph F. Costello, Irmtraud M. Meyer, Connie J. Eaves, Wyeth W. Wasserman, Steven Jones, David Huntsman, Martin Hirst, Carlos Caldas, Marco A. Marra, and Samuel Aparicio. The clonal and mutational

evolution spectrum of primary triple-negative breast cancers. *Nature*, 486:395–399, 6 2012.

[83] Daniel C. Koboldt, Robert S. Fulton, Michael D. McLellan, Heather Schmidt, Joelle Kalicki-Veizer, Joshua F. McMichael, Lucinda L. Fulton, David J. Dooling, Li Ding, Elaine R. Mardis, Richard K. Wilson, Adrian Ally, Miruna Balasundaram, Yaron S.N. Butterfield, Rebecca Carlsen, Candace Carter, Andy Chu, Eric Chuah, Hye Jung E. Chun, Robin J.N. Coope, Noreen Dhalla, Ranabir Guin, Carrie Hirst, Martin Hirst, Robert A. Holt, Darlene Lee, Haiyan I. Li, Michael Mayo, Richard A. Moore, Andrew J. Mungall, Erin Pleasance, A. Gordon Robertson, Jacqueline E. Schein, Arash Shafiei, Payal Sipahimalani, Jared R. Slobodan, Dominik Stoll, Angela Tam, Nina Thiessen, Richard J. Varhol, Natasja Wye, Thomas Zeng, Yongjun Zhao, Inanc Birol, Steven J.M. Jones, Marco A. Marra, Andrew D. Cherniack, Gordon Saksena, Robert C. Onofrio, Nam H. Pho, Scott L. Carter, Steven E. Schumacher, Barbara Tabak, Bryan Hernandez, Jeff Gentry, Huy Nguyen, Andrew Crenshaw, Kristin Ardlie, Rameen Beroukhim, Wendy Winckler, Gad Getz, Stacey B. Gabriel, Matthew Meyerson, Lynda Chin, Raju Kucherlapati, Katherine A. Hoadley, J. Todd Auman, Cheng Fan, Yidi J. Turman, Yan Shi, Ling Li, Michael D. Topal, Xiaping He, Hann Hsiang Chao, Aleix Prat, Grace O. Silva, Michael D. Iglesia, Wei Zhao, Jerry Usary, Jonathan S. Berg, Michael Adams, Jessica Booker, Junyuan Wu, Anisha Gulabani, Tom Bodenheimer, Alan P. Hoyle, Janae V. Simons, Matthew G. Soloway, Lisle E. Mose, Stuart R. Jefferys, Saianand Balu, Joel S. Parker, D. Neil Hayes, Charles M. Perou, Simeen Malik, Swapna Mahurkar, Hui Shen, Daniel J. Weisenberger, Timothy Triche, Phillip H. Lai, Moiz S. Bootwalla, Dennis T. Maglinte, Benjamin P. Berman, David J. Van Den Berg, Stephen B. Baylin, Peter W. Laird, Chad J. Creighton, Lawrence A.

Donehower, Michael Noble, Doug Voet, Nils Gehlenborg, Daniel Di Cara, Juinhua Zhang, Hailei Zhang, Chang Jiun Wu, Spring Yingchun Liu, Michael S. Lawrence, Lihua Zou, Andrey Sivachenko, Pei Lin, Petar Stojanov, Rui Jing, Juok Cho, Raktim Sinha, Richard W. Park, Marc Danie Nazaire, Jim Robinson, Helga Thorvaldsdottir, Jill Mesirov, Peter J. Park, Sheila Reynolds, Richard B. Kreisberg, Brady Bernard, Ryan Bressler, Timo Erkkila, Jake Lin, Vesteinn Thorsson, Wei Zhang, Ilya Shmulevich, Giovanni Ciriello, Nils Weinhold, Nikolaus Schultz, Jianjiong Gao, Ethan Cerami, Benjamin Gross, Anders Jacobsen, Rileen Sinha, B. Arman Aksoy, Yevgeniy Antipin, Boris Reva, Ronglai Shen, Barry S. Taylor, Marc Ladanyi, Chris Sander, Pavana Anur, Paul T. Spellman, Yiling Lu, Wenbin Liu, Roel R.G. Verhaak, Gordon B. Mills, Rehan Akbani, Nianxiang Zhang, Bradley M. Broom, Tod D. Casasent, Chris Wakefield, Anna K. Unruh, Keith Baggerly, Kevin Coombes, John N. Weinstein, David Haussler, Christopher C. Benz, Joshua M. Stuart, Stephen C. Benz, Jingchun Zhu, Christopher C. Szeto, Gary K. Scott, Christina Yau, Evan O. Paull, Daniel Carlin, Christopher Wong, Artem Sokolov, Janita Thusberg, Sean Mooney, Sam Ng, Theodore C. Goldstein, Kyle Ellrott, Mia Grifford, Christopher Wilks, Singer Ma, Brian Craft, Chunhua Yan, Ying Hu, Daoud Meerzaman, Julie M. Gastier-Foster, Jay Bowen, Nilsa C. Ramirez, Aaron D. Black, Robert E. Pyatt, Peter White, Erik J. Zmuda, Jessica Frick, Tara M. Lichtenberg, Robin Brookens, Myra M. George, Mark A. Gerken, Hollie A. Harper, Kristen M. Leraas, Lisa J. Wise, Teresa R. Tabler, Cynthia McAllister, Thomas Barr, Melissa Hart-Kothari, Katie Tarvin, Charles Saller, George Sandusky, Colleen Mitchell, Mary V. Iacocca, Jennifer Brown, Brenda Rabeno, Christine Czerwinski, Nicholas Petrelli, Oleg Dolzhansky, Mikhail Abramov, Olga Voronina, Olga Potapova, Jeffrey R. Marks, Wiktoria M. Suchorska, Dawid Murawa, Witold Kycler, Matthew Ibbs, Konstanty Korski, Arkadiusz Spychała, Paweł Murawa,

Jacek J. Brzeziński, Hanna Perz, Radosław Łaźniak, Marek Teresiak, Honorata Tatka, Ewa Leporowska, Marta Bogusz-Czerniewicz, Julian Malicki, Andrzej Mackiewicz, Maciej Wiznerowicz, Xuan Van Le, Bernard Kohl, Nguyen Viet Tien, Richard Thorp, Nguyen Van Bang, Howard Sussman, Bui Duc Phu, Richard Hajek, Nguyen Phi Hung, Tran Viet The Phuong, Huynh Quyet Thang, Khurram Zaki Khan, Robert Penny, David Mallery, Erin Curley, Candace Shelton, Peggy Yena, James N. Ingle, Fergus J. Couch, Wilma L. Lingle, Tari A. King, Ana Maria Gonzalez-Angulo, Mary D. Dyer, Shuying Liu, Xiaolong Meng, Modesto Patangan, Frederic Waldman, Hubert Stöppler, W. Kimryn Rathmell, Leigh Thorne, Mei Huang, Lori Boice, Ashley Hill, Carl Morrison, Carmelo Gaudioso, Wiam Bshara, Kelly Daily, Sophie C. Egea, Mark D. Pegram, Carmen Gomez-Fernandez, Rajiv Dhir, Rohit Bhargava, Adam Brufsky, Craig D. Shriver, Jeffrey A. Hooke, Jamie Leigh Campbell, Richard J. Mural, Hai Hu, Stella Somiari, Caroline Larson, Brenda Deyarmin, Leonid Kvecher, Albert J. Kovatich, Matthew J. Ellis, Thomas Stricker, Kevin White, Olufunmilayo Olopade, Chunqing Luo, Yaqin Chen, Ron Bose, Li Wei Chang, Andrew H. Beck, Todd Pihl, Mark Jensen, Robert Sfeir, Ari Kahn, Anna Chu, Prachi Kothiyal, Zhining Wang, Eric Snyder, Joan Pontius, Brenda Ayala, Mark Backus, Jessica Walton, Julien Baboud, Dominique Berton, Matthew Nicholls, Deepak Srinivasan, Rohini Raman, Stanley Girshik, Peter Kigonya, Shelley Alonso, Rashmi Sanbhadti, Sean Barletta, David Pot, Margi Sheth, John A. Demchok, Kenna R.Mills Shaw, Liming Yang, Greg Eley, Martin L. Ferguson, Roy W. Tarnuzzer, Jiashan Zhang, Laura A.L. Dillon, Kenneth Buetow, Peter Fielding, Bradley A. Ozenberger, Mark S. Guyer, Heidi J. Sofia, and Jacqueline D. Palchik. Comprehensive molecular portraits of human breast tumours. *Nature 2012 490:7418*, 490:61–70, 9 2012.

[84] Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J.T. Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome Biology*, 20:1–19, 9 2019.

[85] Johanna Wagner, Maria Anna Rapsomaniki, Stéphane Chevrier, Tobias Anzeneder, Claus Langwieder, August Dykgers, Martin Rees, Annette Ramaswamy, Simone Muenst, Savas Deniz Soysal, Andrea Jacobs, Jonas Windhager, Karina Silina, Maries van den Broek, Konstantin Johannes Dedes, Maria Rodríguez Martínez, Walter Paul Weber, and Bernd Bodenmiller. A single-cell atlas of the tumor and immune ecosystem of human breast cancer. *Cell*, 177:1330, 5 2019.

[86] A P Dempster, ; N M Laird, and ; D B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39:1–38, 1977.

[87] Poornima Bhat-Nakshatri, Hongyu Gao, Liu Sheng, Patrick C McGuire, Xiaoling Xuei, Jun Wan, Yunlong Liu, Sandra K Althouse, Austyn Colter, George Sandusky, Anna Maria Storniolo, and Harikrishna Nakshatri. A single-cell atlas of the healthy breast tissues reveals clinically relevant clusters of breast epithelial cells. *Cell Rep. Med.*, 2(3):100219, March 2021.

[88] Tara Chari, Joeyta Banerjee, and Lior Pachter. The specious art of single-cell genomics. *bioRxiv*, page 2021.08.25.457696, 9 2021.

[89] Van Hoan Do and Stefan Canzar. A generalization of t-sne and umap to single-cell multimodal omics. *Genome Biology*, 22:1–9, 12 2021.

[90] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *Nature Communications 2019 10:1*, 10:1–14, 11 2019.

[91] George C. Linderman and Stefan Steinerberger. Clustering with t-sne, provably. *SIAM Journal on Mathematics of Data Science*, 1:313–332, 6 2017.

[92] Benyamin Ghojogh BGHOJOGH, Ali Ghodsi ALIGHODSI, Fakhri Karray KARRAY, and Mark Crowley MCROWLEY. Johnson-lindenstrauss lemma, linear and nonlinear random projections, random fourier features, and random kitchen sinks: Tutorial and survey. *arXiv*, 8 2021.

[93] Alejandro Aguilera-Castrejon, Bernardo Oldak, Tom Shani, Nadir Ghanem, Chen Itzkovich, Sharon Slomovich, Shadi Tarazi, Jonathan Bayerl, Valeriya Chugaeva, Muneef Ayyash, Shahd Ashouokhi, Daoud Sheban, Nir Livnat, Lior Lasman, Sergey Viukov, Mirie Zerbib, Yoseph Addadi, Yoach Rais, Saifeng Cheng, Yonatan Stelzer, Hadas Keren-Shaul, Raanan Shlomo, Rada Massarwa, Noa Novershtern, Itay Maza, and Jacob H. Hanna. Ex utero mouse embryogenesis from pre-gastrulation to late organogenesis. *Nature 2021 593:7857*, 593:119–124, 3 2021.

[94] Daniela F. Quail and Johanna A. Joyce. Microenvironmental regulation of tumor progression and metastasis. *Nature medicine*, 19:1423, 11 2013.

[95] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods*, 14(9):865–868, September 2017.

[96] Ricard Argelaguet, Anna S E Cuomo, Oliver Stegle, and John C Marioni. Computational principles and challenges in single-cell data integration. *Nat. Biotechnol.*, 39(10):1202–1215, October 2021.

[97] Yang Xu and Rachel Patton McCord. Diagonal integration of multimodal single-cell data: potential pitfalls and paths forward. *Nat. Commun.*, 13(1):3505, June 2022.

[98] Mika Sarkin Jain, Krzysztof Polanski, Cecilia Dominguez Conde, Xi Chen, Jongeun Park, Lira Mamanova, Andrew Knights, Rachel A Botting, Emily Stephenson, Muzlifah Haniffa, Austen Lamacraft, Mirjana Efremova, and Sarah A Teichmann. MultiMAP: dimensionality reduction and integration of multimodal data. *Genome Biol.*, 22(1):346, December 2021.

[99] Yuhan Hao, Stephanie Hao, Erica Andersen-Nissen, William M Mauck, 3rd, Shiwei Zheng, Andrew Butler, Maddie J Lee, Aaron J Wilk, Charlotte Darby, Michael Zager, Paul Hoffman, Marlon Stoeckius, Efthymia Papalexi, Eleni P Mimitou, Jaison Jain, Avi Srivastava, Tim Stuart, Lamar M Fleming, Bertrand Yeung, Angela J Rogers, Juliana M McElrath, Catherine A Blish, Raphael Gottardo, Peter Smibert, and Rahul Satija. Integrated analysis of multimodal single-cell data. *Cell*, 184(13):3573–3587.e29, June 2021.

[100] Brodmann K. *Brodmann's: Localisation in the Cerebral Cortex*. Springer, 2010.

[101] WolframMathWorld. Manifold, 2022. `https://mathworld.wolfram.com/Manifold.html`, Last accessed on 2022-12-23.

[102] Ricard Argelaguet, Damien Arnol, Danila Bredikhin, Yonatan Deloro, Britta Velten, John C Marioni, and Oliver Stegle. MOFA+: a statistical framework for comprehensive integration of multi-modal single-cell data. *Genome Biol.*, 21(1):111, May 2020.

[103] Yuge Shi, N. Siddharth, Brooks Paige, and Philip H. S. Torr. Variational mixture-of-experts autoencoders for multi-modal deep generative models, 2019.

[104] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[105] Emile Mathieu, Tom Rainforth, N. Siddharth, and Yee Whye Teh. Disentangling disentanglement in variational autoencoders, 2018.

[106] Mike Wu and Noah Goodman. Multimodal generative models for scalable weakly-supervised learning. *arXiv*, 2018.

[107] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

[108] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[109] Edward C. Klatt. Gastrointestinal pathology, 2022. `https://webpath.med.utah.edu/GIHTML/GI112.html`, Last accessed on 2022-12-11.

[110] Irene Y. Chen and Diana Agostini-Vulaj. Hepatocellular carcinoma, 2022. `https://www.pathologyoutlines.com/topic/livertumorhcc.html`, Last accessed on 2022-12-11.

[111] Chandler D Gatenbee, Ann-Marie Baker, Ryan O Schenck, Maximilian Strobl, Jeffrey West, Margarida P Neves, Sara Yakub Hasan, Eszter Lakatos, Pierre Martinez, William C H Cross, Marnix Jansen, Manuel Rodriguez-Justo, Christopher J Whelan, Andrea Sottoriva, Simon Leedham, Mark Robertson-Tessi, Trevor A Graham, and Alexander R A Anderson. Immunosuppressive niche engineering at the onset of human colorectal cancer. *Nat. Commun.*, 13(1):1798, April 2022.

Appendices

A

ADAM17	FGFR1
ADAM	GBE1
AK3	HEGF
ALDOA	HK2
ANGPTL4	KDM3A
APOBEC3A	KRT7
APOBEC3B	KRT19
APOBEC3C	LDHA
APOBEC3D	LDLR
APOBEC3F	MIF
APOBEC3G	MKI67
APOBEC3H	MTFR1
AREG	MUC1
BNIP3	NDRG1
CA9	P4HA1
CDH1	PFKP
CHCHD2	PGK1
CXCL12	PYGL
DDIT4	SF3B5
EGFR	SLC16A1
ENO1	SLC16A3
ERBB2	SLC2A1
ERBB3	SLC6A8
EREG	TFRC
ERO1A	TGFA
ESR1	TGFBR1
FGF13	TPI1
FGF7	VEGFA

Table A.1: HUGO gene names for the CARTANA hypoxia-immune bespoke panel.